INFORMATION - ENTROPY CONCEPTS FOR NUTRITIONAL SYSTEMS

Dissertation for the Degree of Ph. D. MICHIGAN STATE UNIVERSITY JEROME PAUL HARPER 1976





This is to certify that the

thesis entitled

Information-Entropy Concepts for Nutritional Systems

presented by

Jerome Paul Harper

has been accepted towards fulfillment of the requirements for

Ph. D. degree in Agricultural Engineering

Serisham

Major professor

Date aug. 5, 1976

**O**-7639







#### ABSTRACT

#### INFORMATION-ENTROPY CONCEPTS FOR NUTRITIONAL SYSTEMS

By

### Jerome Paul Harper

The objective of this dissertation is to view nutritive processes as communication systems for transmitting dietary nutritional information. This study utilizes information theory concepts in the analysis of nutritive communication systems. The concept of information-entropy is used to derive the information capacity of the system's dietary inputs and metabolic requirements.

The major process investigated is the system transmitting information as amino acids for protein metabolism. First, a gene-protein channel is defined and hpothesized to be the determinant of the metabolic informationentropy requirements for amino acids. A nutritive communication system is then postulated which contains five basic components: (1) information source (food protein), (2) encoder (intestinal amino acid transport system), (3) channel (circulatory system), (4) decoder cellular amino acid transport system), and (5) receiver

Jerome Paul Harper

(cellular amino acid pool). The transmission capacity of amino acid information depends upon cellular metabolic requirements which control the decoding capacity, and thus the overall transmission efficiency. Cost of transmission is defined as the ability of an information source to satisfy metabolic requirements during a fixed time period. A familiar rank-frequency distribution of information theory, Zipf's law, is employed to order source proteins on the basis of metabolic cost. Net protein value is shown to be proportional to the inverse protein rank (quality). A single channel model yields a protein ranking similar to chemical score, while the multichannel model generates a ranking similar to Oser's essential amino acid index. The multichannel model could be adapted to consider amino acid catabolism by the liver (an important loss of information in the channel), and predict a new protein ranking termed the "essential amino acid retention index."

The other study concerns the information-entropy of carbohydrate polymers. The hydrolysis of these polymers is regarded as a metabolic encoding process. The dietary carbohydrate message has to be reduced to the monomer or dimer form if it is to be transmitted through the nutritional channel (i.e., circulatory system). The cost of encoding (time/monomer) is equated with the inverse activity of enzymatic hydrolysis (monomers/time). The ranking of carbohydrate message length (degree of polymerization) with respect to the rate of encoding (hydrolytic activity) is shown to be identical to the ordering scheme dictated by Zipf's law.

Approved

Approved

### INFORMATION-ENTROPY CONCEPTS

### FOR NUTRITIONAL SYSTEMS

Ву

Jerome Paul Harper

### A DISSERTATION

Submitted to Michigan State University in partial fulfillment of the requirements for the degree of

### DOCTOR OF PHILOSOPHY

Department of Agricultural Engineering

To Christine

### ACKNOWLEDGMENTS

I wish to thank my major professor Dr. J. B. Gerrish for his guidance and assistance on this dissertation and during my doctoral candidacy.

Also, I wish to thank the members of my committee, Dr. M. Z. v. Krzywoblocki, Dr. J. W. Thomas, Dr. D. K. Anderson, and Dr. J. B. Holtman for their thoughts and efforts in the preparation of my dissertation. In particular, I want to acknowledge Dr. Krzywoblocki, "Ziggy" to his friends and students, for his tutelage in the field of information theory.

iii

### TABLE OF CONTENTS

															Page
LIST OF	TABLE	ES	•	•	•	•	•	•	•	•	•	•	•	•	vi
LIST OF	FIGUI	RES	•	•	•	•	•	•	•	•	•		•	•	ix
LIST OF	TERMS	5	•	•	•	·	•	·	·	•	•	•	•	•	x
Chapter															
I.	INTRO	DUC	TIO	N	•	•	•	•	•	•	•	•	•	•	1
II.	LITER	RATU	RE	REV	IEW	AN	DF	ERS	PEC	TIV	Е	•	•	•	5
	2.1	Som	ет	her	mod	yna	mic	As	pec	ts	of				
		Ent	rop	У			•			•			•	•	5
	2.2	Ent	rop	y a	nd	Inf	orm	nati	on	The	ory		•	•	17
	2.3	Ent	rop	y,	Inf	orm	ati	on	and	Bi	olc	gy	•	•	33
	2.4	Ent	rop	y,	Inf	orm	ati	on	and	Nu	tri	tic	n	•	39
III.	INFOR	TAMS	ION	AN	DT	HE	QUA	LII	Y C	FF	ROT	EIN	IS	•	44
	3.1	Ind An	ice Inf	s o	f P ati	rot	ein Ent	Qu	ali	ty	i	f	•	•	44
	3 3	Pro	tei	n Q	ual	ity Inf	orm			Fnt	, ron	•	•	•	54
	5.5	App	roa	ch	•	•	•	•	•	•	•	•	•	•	77
IV.	INFOR	RMAT DHYD	ION RAT	AN E P	D T OLY	HE MER	HYC S	ROL		s c	·			•	94
	4.1	Asp Hyd	ect: rol	s o ysi	f C s a	arb nd	ohy Met	dra abo	te lis	Str m	uct	ure •	:		94
	4.2	An	Enc	odi	ng	Mod	el	for	Ca	rbc	hyd	rat	e		
	4 3	Acc	Orm	mon	on t o	f +	ho	Car	hoh	vdr	·	•	•	•	98
	4.5	Inf	orm	ati	on-	Ent	rop	y A	nal	ysi	s	•	•	•	107
v.	DISCU	JSSI	ON	•	•	•	•		•	•	•	•	•	•	122
	5.1	Nit Ent	roge	en Y	Ret •	ent •	ion •	an •	d I •	nfc •	rma •	tic •	n- •		122

.

,

Chapter															Page
	5.2	Th Pr In	e I ote for	nfc in mat	Meta	io bo	n-E lis	m:	copy Sur	Mo	del ry	fo	r	•	133
	5.5	An	Ap	pra	isal			•	•	•	•	•	••	•	135
VI.	CONC	LUS	ION	s		•		•	•	•	•	•			138
REFEREN	CES														139

## LIST OF TABLES

Table		Page
3.1.1	Listings of Biological Value, Net Pro- tein Utilization, Net Protein Value, and Protein Efficiency Ratio Scores with their Respective Rankings (Source: FAO)	49
3.1.2	Matrix of Correlation Coefficients Relating Biological Value, Net Protein Utilization, Net Protein Value, and Protein Efficiency Ratio Scores	50
3.1.3	Matrix of Correlation Coefficients Relating Biological Value, Net Protein Utilization, Net Protein Value, and Protein Efficiency Ratio Ranks $(\rho_s)$ .	50
3.3.1	Amino Acid Content (micromoles per gram N) of Food Proteins (Source: Eggum)	80
3.3.2	Amino Acid Content (micromoles per gram N) of Food Proteins (Source: FAO)	81
3.3.3	Biological Values and Net Protein Values of Sixteen Test Diets of Rats and Baby Pigs (Source: Eggum)	83
3.3.4	Information-Entropy Indices for Sixteen Different Food Proteins	84
3.3.5	Matrix of Correlation Coefficients for Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on amino acid content of dietary protein)	85
3.3.6	Matrix of Spearman's Rank Correlation Coefficients for Ranks of Information- Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on amino acid content of dietary protein)	86

# Table

3.3.7	Matrix of Correlation Coefficients for Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on available amino acid content of dietary protein)	•	88
3.3.8	Matrix of Spearman's Rank Correlation Coefficients for Ranks of Information- Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on avilable amino acid content of dietary protein)	•	88
3.3.9	Matrix of Correlation Coefficients for Zipfian (log-log) Analysis of Information-Entropy Model	•	90
3.3.10	Matrix of Slopes of Zipfian (log-log) Analysis of Information-Entropy Model	•	91
4.3.1	Degree of Polymerization and Enzyme Kinetic Data of Amylose	•	111
4.3.2	Degree of Polymerization and Activity Data of Cellulose, with activity in (moles/sec.) x $10^{-9}$	•	111
4.3.3	Correlation and Regression Analysis of Activity Data Versus Degree of Polymerization	•	112
4.3.4	Hydrolysis of Amylose Polymers with $\beta$ -Amylase, and Degree of Polymerization	•	115
4.3.5	Correlation and Regression Analysis of Hydrolysis and Degree of Polymeriza- tion	•	115
4.3.6	Chain Length Fractionalization of a Polydisperse Carbohydrate System as a Function of Its Degradation	•	119
3.3.7	Chain Length Behavior of Polydisperse Carbohydrate Systems	•	120

.



# Table

5.1.1	Correlation Coefficients Among Essential Amino Acid Retention Indices and Experimental Protein Values of Rats and Pigs
5.1.2	Linear Regression Coefficients Among the Essential Amino Acid Retention Indices and Experimental Net Protein Values of Rats and Pigs
5.1.3	Slopes for Regression Analysis Among Essential Amino Acid Indices and Experi- mental Net Protein Values of Rats and Pigs

### LIST OF FIGURES

### Figure

3.2.1	Graphical Representation of Transcrip- tion of Genetic Information in Protein	
	Synthesis	56
3.2.2	Idealized Communication System for Transmission of Genetic Information	59
3.2.3	Idealized Communication System for the Transmission of Amino Acid Informational Molecules	64
3.3.1	$\bar{H}^0_X(\text{EAA}),$ the Average Information-Entropy, Versus Zipfian-Rank-Ordering for Net Protein Value for Rats	92
4.3.1	The Degree of Polymerization Versus Enzymatic Activity for $\beta\text{-Amylase}$	.09
4.3.2	The Degree of Polymerization Versus Enzymatic Activity for Cellulose <u>A</u> ( <u>Penicillium Notatum</u> )	.10
5.1.1	Graph of Information-Entropy Indices $I_{\rm x}^0({\rm EEAR}_{\rm e})$ and $I_{\rm x}^0({\rm EEAR}_{\rm e})$ Versus Net Protein Value for Rats (Source: Eggum) 1	31
5.1.2	Graph of Information-Entropy Indices $I_{X}^{O}\left(\text{EAAR}_{e}\right)$ and $I_{X}^{O}\left(\text{EAA}_{e}\right)$ Versus Net Protein Value for Pigs (Source: Eggum) 1	32



## LIST OF TERMS

A	Canonical normalization constant
A <sub>1</sub> ; A <sub>2</sub>	Constants
a <sub>1</sub> ; a <sub>2</sub>	Total enzyme activity
alj	Enzymatic activity for reaction with one substrate bond
aa <sub>j</sub>	j <sup>th</sup> amino acid
am. J	Maximum j <sup>th</sup> amino acid frequency, magnitude of amino acid variable at receiver
ar xj	Absolute ranking of j <sup>th</sup> amino acid
ax; as j j	Frequency of j <sup>th</sup> amino acid in proteins x and s
В	Body nitrogen of protein-fed animals
В'	Constant
b; b'	Constants
<sup>B</sup> k	Body nitrogen of non-protein-fed animals
BV	Biological value
BV(x <sub>j</sub> )	Biological value of j <sup>th</sup> amino acid
с	Channel capacity of a system
C <sub>I</sub> ; C <sub>J</sub>	Total costs of message
° <sub>i</sub> ; ° <sub>j</sub>	Cost of i <sup>th</sup> or j <sup>th</sup> symbol or word
C <sub>xj</sub> ; C <sub>mj</sub>	Capacity of j <sup>th</sup> channel at source and receiver, respectively

х



c <sub>xj</sub> ; c <sub>mj</sub>	Cost of source aa. and receiver aa <sub>j</sub> , respectively <sup>j</sup>
с <sub>N</sub>	Total cost of message
CS	Chemical score
c <sup>x</sup> ; c <sup>s</sup>	Concentration of protein in diet
D	Digestibility
D <sub>i</sub>	Hydrolysis coefficient
$D(s_j); D(x_j)$	Digestibility of protein
DP	Degree of polymerization
DP j	Length of message
d(x,t)	Distribution function
e	Energy unit
e-subscript	Denote Eggum as data source
E	Macroscopic energy
EAA	Essential amino acid set
EAAI	Essential amino acid index
Ej	Replica energy in canonical ensemble
EF j	Encoding efficiency
F	Fecal nitrogen
f-subscript	Denotes FAO as data source
f <sub>c</sub>	Catabolism factor
F <sub>k</sub>	Endogenous fecal nitrogen
fo	Conservation factor
Н	Boltzman H-function
<sup>h</sup> j	Information of j <sup>th</sup> symbol



Hmax	Maximum of H-function
H <sub>n</sub>	n <sup>th</sup> order multivariate information capacity
$H_n^I$	Independent n <sup>th</sup> order multivariate information capacity
H <sup>I</sup> <sub>2</sub>	Independent bivariate information capacity
<sup>H</sup> 2	Bivariate information capacity
H <sub>r</sub> (aa <sub>j</sub> )	Entropy of receiver
H(x)	Information-entropy content of $\underline{x}$
H <sub>x</sub> (aa <sub>.</sub> )	Characteristic information-entropy of protein
H <sub>x</sub> (aaj)	Total information-entropy of j <sup>th</sup> variable
$     \overline{H}_{x}^{(EAA)}; $ $     \overline{H}_{x}^{0}^{(EAA)}; $	Average information-entropy of essential amino acid set
н <sub>s</sub> (ЕАА)	
I(aaj)	Channel transmission rate at source
ıj	j <sup>th</sup> channel transmission rate
I <sub>s</sub>	Stored information
I <sub>s1</sub> ; I <sub>s2</sub>	Univariate and bivariate divergence
I <sub>x</sub> (CS)	Information index
I (EAAR); I <sup>0</sup> (EAAR) X	Essential amino acid retention index
I <sub>x</sub> (NPV) ; I <sup>0</sup> <sub>x</sub> (NPV)	Information-entropy indices
k	Constant for converting ln to $\log_2$
<sup>k</sup> b	Boltzman's constant

xii



к <sub>т</sub>	Michaelis-Menten constant
m(DP <sub>j</sub> )	Frequency of messages with length DP j
min H <sub>s</sub> (aa <sub>j</sub> )	Log frequency of I <sub>x</sub> (CS) for stan- dard protein
min H <sub>x</sub> (aa <sub>j</sub> )	Log frequency of $I_x(CS)$ for protein <u>x</u>
μ <sub>k</sub>	Chemical potential of $j^{th}$ particle
N	Total number of particles or words
N(cj)	Total number of words of cost $c_j$
Ng	Total number of glucose units in source
<sup>N</sup> gi <sup>, N</sup> gj	Number of glucose monomers in system
N <sub>I</sub>	Nitrogen intake
<sup>N</sup> Ik	Nitrogen intake of non-protein-fed animals
NPU	Net protein utility
NPV	Net protein value
NPV(x <sub>j</sub> ); NPV(s <sub>j</sub> )	Net protein value of j <sup>th</sup> amino acid
NPV <sub>x</sub> (EAA) ; NPV <sub>s</sub> (EAA)	Log-averages of net protein values
<sup>n</sup> j	Number of $j^{th}$ particles
Р	Permutability factor
P <sub>c</sub>	Canonical probability function
PER	Protein efficiency ratio
Pgc	Grand canonical probability function

<u>1</u>



P(i); P(j)	Probability of i <sup>th</sup> or j <sup>th</sup> element
P(ij)	Joint probability of $i^{th}$ and $j^{th}$ elements
Pj	Probability of j <sup>th</sup> state
P(j/i)	Conditional probability of j <sup>th</sup> state
Pmc	Microcanonical probability function
đ	Heat
R	Redundancy
ρ <sub>s</sub>	Spearman's rho
<sup>r</sup> xj <sup>; r</sup> sj	Relative rank of j <sup>th</sup> amino acid in protein $\underline{x}$
r <sub>x</sub>	Rank of protein $\underline{\mathbf{x}}$
S	Entropy
s'	Substrate concentration
s <sub>m</sub>	Statistical entropy of macrosystem
т	Absolute temperature
t	Time
U	Urinary nitrogen
U <sub>k</sub>	Endogenous urinary nitrogen
v	Reaction velocity
v <sub>m</sub>	Maximum reaction velocity
Z	Partition function



### CHAPTER I

### INTRODUCTION

The title of this dissertation contains a compound word, "information-entropy." This word carries with it two different and distinct concepts which together represent somewhat of a union of information theory and thermodynamics. This union can be viewed as generalizing the study of thermodynamics.

My notion of a generalized theory of thermodynamics revolves about a simple hypothesis on the occurrence of events. All events occur more or less frequently for two reasons: (a) There exist physical reasons favoring certain states, and (b) there exist some mental reasons favoring certain states. Therefore, if one is to interpret phenomena, a theory, methodology, or principle is needed to quantify the frequency of physical phenomena, and understand what quantification means.

In science, the quantification of observable phenomena is accomplished by calculating the entropy of the system. In fact, the Second Law of Thermodynamics implies that systems which cannot be quantified because



their behavior is so random have no utility. It was the purpose of men such as Boltzman, with his permutability factor, and Gibbs, with his interpretation of the behavior of volume in phase space, to develop within the science of thermodynamics the ability to quantify nature.

Given that entropy is a measure which strives to quantify, how does it relate to information theory? Information theory is the science of quantification. Using its concepts and applying its theorems allows us to understand how we are quantifying a system. This is why I will later stress the importance of Zipf's law, an empirical rank-size rule, in its information theory context of frequency-cost relationships. The empirical observation of Zipfian behavior is of little benefit if one does not recognize that such behavior depicts a system's organization. It is, then, no mere coincidence that thermodynamic measures of entropy and the communication measure of information are similar.

Because it is important to quantify phenomena, I have sought in my thesis to develop an informationentropy methodology for analyzing nutritional systems. Frequency, pattern, and organization are important concepts in nutrition and a proper format should be developed for quantifying them. My conceptualization of entropy in thermodynamics is that of quantification. Little difference between the meanings of "entropy" in



thermodynamics, and "information" in information theory exists. Both strive to achieve the same end, quantification, which is the deeper meaning of entropy.

The quantification of nutritional systems is begun by assuming the organism under study exists as a biological information processing system. The information being processed here is nutritional information. The source of nutritional information for the organism is the diet, which contains a vast array of different nutritional signals, each signal varying in its frequency of occurrence and each diet providing a distinctive pattern of signals. This information is then fed into a highly organized biological communication system which distributes and integrates the nutritional information to provide the necessary chemical order (nutrients) for the continuance of the organism's metabolic processes.

The nutritive system which best fits into the above sequence of events is the protein metabolic system. An extensive study of this system will be presented in the text, and the relationships between the frequencies of nutritional information (amino acids) and different nutritional frequency patterns of proteins can be used as a measure of the protein quality of the diet.

The carbohydrate study measures the information or entropy of macromolecules. The amount of information which they possess is based upon the frequency of

nutritional signals (glucose units) per molecule. The ability of the carbohydrate message to be interpreted (hydrolyzed) by the organism's enzymes depends on message length.

The concept of "information-entropy" as used in this study is thus defined as "the frequency of a nutritional event" (e.g., the occurrence of an amino acid in the diet or glucose unit in a polymer). This nutritional frequency is then shown to favor a particular metabolic state.


in the second second

## CHAPTER II

## LITERATURE REVIEW AND PERSPECTIVE

The relationship between entropy, information and biology is complex. The following chapter presents classical and statistical mechanical ideas on entropy, their relationships to information, and the role of information-entropy in biological systems. This chapter is partically an historical review and partially a commentary on the subject. Its purpose is to provide the reader with a perspective on both subjects' scientific aspects and my own conceptualization of the interrelationships among entropy, information and biology.

## 2.1 Some Thermodynamic Aspects of Entropy

Perhaps the best way to present the concept of information-entropy is to begin with the development of the entropy principle in thermodynamics. Early ideas on heat were based on the study of steam engines, and Sadi Carnot's (1) notes in 1824 on the efficiencies of these engines are often regarded as the starting point of thermodynamics. The concept of entropy was intimately associated with views on the nature of energy, which was



thought to possess one of two qualities: (1) to be free and available to do mechanical work, or (2) to be bound and incapable of mechanical work. A qualitative degradation of energy from the free to the bound form was invariably observed to occur, and several rules were formulated to describe this phenomenon.

Clausius (2) stated, "Heat can never, of itself, flow from a colder to a hotter temperature." Thomson's position, on the other hand, was, "It is impossible to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest surroundings" (3).

These statements are similar, and the principle of physics which was derived from them is known as the Second Law of Thermodynamics. Clausius presented in 1865 the classic formulation of this law:

The entropy of the universe at all times moves toward a maximum.

The sense of entropy here is like the notion of bound or latent energy with the added constraint of being quantified at a particular temperature. The following equation shows the relationship (4):

$$Entropy = \frac{Bound energy}{Absolute temperature} .$$
(2.1.1)

A refinement of the above relationship was obtained from analysis of the behavior of an elementary

heat engine in the Carnot cycle. It was reasoned that if heat, q, were allowed in or out of the cycle only in infinitesimally small increments, then q could be approximated by its differential form, dq. Utilizing this differential form the Carnot cycle heat behavior can be described by the following integral form:

$$\oint \frac{dq}{T} = 0 \text{ (reversible)}, \qquad (2.1.2)$$

where T is the absolute temperature. The above is an interesting mathematical form, for it is an equation which implies that one has uncovered an exact differential measure, another state variable, different from energy, which describes a system's thermodynamic behavior as the state changes from <u>a</u> to <u>b</u>. The new state function was entropy, S, and was formally defined:

 $dS = \frac{dq}{T}$  (for a reversible process) (2.1.3)

A mathematical approach for the derviation of caloric entropy proposed by Caratheodory (5) verified that dq/T is an exact differential only if the process is reversible.

Clausius used the efficiency concept to illustrate a general difference in the entropy behavior between reversible and irreversible processes. His result, known as the Inequality of Clausius, states that the efficiency of a reversible cyclic process, like the Carnot cycle,

is always greater than that of an irreversible cycle. Clausius' comprehension of the fact that all real (i.e., irreversible) processes result in an entropy increase of their surroundings, led to his statement of the Second Law of Thermodynamics, which was the climactic effort of the classicists in heat theory.

Because of the macroscopic nature of the Carnot cycle study, it is not amenable to mechanistic analysis. Such analysis requires a more detailed description of the phenomenon, a microscopic picture, so that the macroscopic parameters such as temperature, pressure, etc., are understood as the aggregated mechanical behavior of the elementary masses (atoms, molecules) in the system. In a mechanical model, the possibility exists for defining the state variable, entropy, irrespective of the process being reversible or in equilibrium. The groundwork for such a model began in the late nineteenth century and was to be the basis for understanding the formulation of entropy.

The development of a mechanical explanation of the Second Law of Thermodynamics can primarily be attributed to Ludwig Boltzman. His explanation of the Second Law has become the mainstay of statistical mechanics. Boltzman began his studies in 1866 (6) which were highlighted in 1872 (7) with the publication of a long memoir which gave the first derivation of the

irreversible increase of entropy based on the laws of mechanics and also upon those of probability. In this memoir Boltzman presented a mathematical proof of the second thermodynamic law by illustrating the uniqueness of Maxwell's velocity distribution law (8) as a descriptor of the equilibrium state. Maxwell had shown for gases that his distribution was stationary and Boltzman expanded the application of this proof by demonstrating that whatever the initial state of a gas, it approaches a limit in the distribution of Maxwell.

In this proof Boltzman derived a partial differential equation for a distribution function d(x,t), with respect to time, the distribution function representing the number of molecules per unit volume with kinetic energies at time <u>t</u> lying within an interval of <u>x</u> to (x + dx). He showed that Maxwell's function is stationary and makes  $\partial d(x,t)/\partial t$  vanish. The next phase called for the introduction of an auxiliary function called H, defined:

$$H = \int_0^\infty d(x,t) \{ \ln [d(x,t)/\sqrt{x}] - 1 \} dx, \qquad (2.1.4)$$

which he proved can only decrease with time due to the symmetry characteristics of the collision process and the possibility of inverse collisions. It was then shown that the Maxwell's distribution function minimizes



the <u>H</u> function, proving that regardless of the initial distribution of d(x,t) the final or equilibrium state is realized in the Maxwell distribution. Even more important than this fact, Boltzman pointed out, was that the quantity H was proportional (with a negative proportionality constant) to the entropy of the gas.

Needless to say, this result caused a considerable degree of interest and before long criticisms of his approach arose. It was Boltzman's response (9) to one of his critics that resulted in the formulation of the second thermodynamic law as an expression of the laws of probability. He showed that the entropy of a state is reflected by its probability and an increase in entropy merely reflects a shift from less to more probable states. He employed a discrete model to illustrate this probabilistic nature of entropy. It was hypothesized that a collection of N particles possessed energies which were integral multiples, <u>j</u>, of a basic energy unit, <u>e</u>. The number of particles having the j x e energy is denoted by  $n_j$ , such that the sum over <u>j</u> of  $n_j$  equals <u>N</u>.

For a complete assessment of this system of particles, a listing of all the individual molecular energies would be required. To attain this assessment a permutability measure,  $\underline{P}$ , the number of different arrangements (microstates) for a given distribution was constructed by the equation:



$$P = \frac{N!}{n_0! n_1! \dots n_j!} . \qquad (2.1.5)$$

Boltzman then reasoned that the most probable distribution was the one where  $\underline{P}$  is maximized. To find the maximum for  $\underline{P}$  he first used Stirling's approximation for factorials (10) and proceeded to deduce the following equality:

$$\ln P = -\sum_{j} n_{j} \ln n_{j} + \text{constant} . \qquad (2.1.6)$$

Recall equation (2.1.4) and recognize the similarities between  $n_j$  and d(x,t) and that the negative of the <u>H</u> function is entropy. The beauty of Boltzman's proof becomes readily apparent: he has, first, found the distribution which maximizes <u>P</u>; second, shown the relationship between <u>P</u> and entropy; and third, knows the  $n_j$ 's will have a Maxwell distribution when the entropy of the system is maximum. The classic formulation Boltzman gave for statistical entropy,  $S_m$ , of the macrosystem in terms of its microstate distribution is:

$$S_m = k_b \cdot \ln P \tag{2.1.7}$$

where k<sub>b</sub>is known as Boltzman's constant which is determined by dividing the gas constant by Avogadro's number.

The notion of probability is ascertained from  $\underline{P}$ , the permutability factor. Consider the logarithmic expansion of equation (2.1.5) and apply Stirling's approximation; the result has the form:



1.

catalogus

$$\ln P \cong N (\ln N-1) - \sum_{j} n_{j} (\ln n_{j} - 1)$$

$$= N \ln N - \sum_{j} n_{j} \ln j_{n}$$

$$= -N \sum_{j} (n_{j}/N) \ln n_{j}/N . \qquad (2.1.8)$$

The quantity  $n_j/N$  is identical to the probability of the j<sup>th</sup> microstate which shall be denoted as P<sub>j</sub>. Substituting equation (2.1.8) into equation (2.1.7) for the macrosystem entropy:

$$S_{m} = -k_{b}N \sum_{j} P_{j} \ln P_{j}$$
 (2.1.9)

The behavior of this function is identical to that in equation (2.1.3), and generates the most probable distributions when P is maximized.

Because one is greatly limited in his knowledge of the microsystem structure, another successful approach which overcame such limitations appeared at the beginning of the twentieth century. It was developed by Gibbs (11). The Gibbsian approach was proposed to show how microscopic "behavior" determined the total thermodynamic picture. Primarily, this was attained by employing an abstraction which Gibbs called an ensemble. In essence, this was a statistical-mechanical theory which could be generalized as a statistical theory of systems of differential equations (12).



An ensemble can be defined as a collection of a large number of identical replicas of the representative system. These replicas are all independently performing the identical irreversible process. The main assumption of ensemble theory is that the instantaneous macroscopic state is related to the average of the replicas' states taken over all the replicas.

The ensemble's macroscopic conditions can dictate the probability distributions for each replica by affecting energy and motion. Then, for different macroscopic conditions, different ensemble types can be identified. The following are the three most common ensembles employed (13):

Microcanonical ensemble: A statistical ensemble of closed, energetically isolated systems in a constant volume, a replica here can be thought of as being enclosed in an adiabatic shell where neither exchange of energy nor of particles is allowed. The rigidity of the constraints implies that a simple probability function holds for this type of ensemble. The microcanonical probability function, P<sub>mc</sub>, is constant and of the form

 $P_{\rm MC} = 1/P$  . (2)

The Maxwell-Boltzman distribution can be used to calculate the probability function of a microcanonical ensemble.

<u>Canonical ensemble</u>: A statistical ensemble in thermal contact with a thermostat, here the replica is permitted to exchange energy with another system whose energy is so large by comparison that its state remains unchanged. The canonical probability function,  $P_C$ , is therefore a function of the energy  $E_j$  of each replica and its form is exponential:

 $P_{C} = A \exp(-B'E_{j})$ .

(2.1.11)

(2.1.10)

A is a constant fixed by normalizaton, <u>B</u>' is the inverse of the product of the Boltzman constant and the absolute temperature. The term  $\exp(-B^{*}E_{j})$  is called the Boltzman factor.

<u>Grand canonical ensemble:</u> A statistical ensemble which can exchange both energy and particles with its surroundings, such an ensemble can be conceived of as a box in contact with a thermostat and possessing permeable walls. The grand canonical probability function,  $P_{\rm qc}$ , is based both on considerations of energy and particles:

$$\begin{split} P_{gc} &= A \, \exp{(-B'E_j - B' \, \sum\limits_{j}^{n} n_j \mu_k)} \quad (2.1.12) \\ \text{where } \mu_k \text{ is the chemical potential of the } j^{th} \\ \text{particle type.} \end{split}$$

After deciding what ensemble to employ, the remaining problem in the Gibbs approach is that of computing what is known as the partition function. The partition function is probably the most important concept in statistical mechanics today, from which important thermodynamic variables (including entropy) can be estimated. The partition function is a very simple mathematical form that depicts the distribution or partitioning of the system among the various energy levels or quantum states. To calculate, sum the Boltzman factors for all the different states (14):

$$Z = \sum_{j} \exp(-B'E_{j}) . \qquad (2.1.13)$$

The partition function has an important statistical relationship to the probability of the system:

 $P_{j} = n_{j}/N = \exp(-B'E_{j})/Z$  (2.1.14)

which gives the function immense utility in thermodynamic calculations.

An expression for entropy in terms of the partition function can be readily deduced. Recall that the classical definition for entropy is the reversible differential energy change divided by the absolute temperature. As the energy changes in the process so will the partition function with respect to B' and  $E_4(15)$ :

$$d \ln z = -E dB' - (B'/N) \sum_{j} n_{j} dE_{j}$$
 (2.1.15)

(ln Z is preferred to Z because of its additive properties).

By performing a Legendre transformation on equation (2.1.15) and collecting terms, we cause the differential energy or heat change of the system to become:

 $dE = T \cdot dS = d(\ln Z + B'E)/B'$  (2.1.16)

or, alternatively, the entropy is:

 $dS = k_{\rm b} \cdot d(\ln Z) + d(E/T)$  (2.1.17)

or

$$S = k_{\rm b} \cdot \ln Z + E/T$$
. (2.1.18)

Equation (2.1.17) can be converted to the Boltzman equation for entropy after we recognize two relationships:

$$E = \sum_{j} P_{j}E_{j}$$
(2.1.19)

$$-B'E_{j} = \ln(P_{j}Z) . \qquad (2.1.20)$$

Equation 2.1.19 states that the macroscopic energy of the system is the expected value determined from the energies of each microstate, and equation (2.1.20) is the logarithmic form of equation (2.1.14). Substituting into equation (2.1.18) we have:

$$S = k_{b} \ln z + k \sum_{j} P_{j}(B'E_{j})$$
$$= k_{b} \ln z - k \sum_{j} P_{j} \ln P_{j} - k \sum_{j} P_{j} \ln z$$
$$= k_{b} \sum_{j} P_{j} \ln P_{j} . \qquad (2.1.21)$$

The above equation is identical to equation (2.1.9), the Boltzman entropy, divided by N.

In spite of the obvious similarities between the Gibbsian approach and that of Boltzman, there are also pertinent differences. Boltzman's entropy is a measure which does not consider interparticle forces, and thus neglects the effects of potential energy and the effect of interparticle forces on pressure; Gibbs' entropy takes into account all the energy and total pressure (16). The question then arises, what is the true thermodynamic entropy of a system?

The current thought on this question tends to the viewpoint that true thermodynamic entropy is difficult

16

and

to define because the partitioning or experimental conditions of the system depend upon the human element (17). This "anthropomorphic" aspect of entropy imparts a considerable degree of arbitrariness, making a definition of true thermodynamic entropy essentially impossible. However, it should be remembered that irrespective of the manner in which the system's partitioning is accomplished, the partition-dependent behavior is not arbitrary but follows a course dictated by the Second Law. Consequently, when studying the entropic behavior of a system the most difficult problem is to state what questions we want to resolve and to formulate entropic measures which allow their resolution.

## 2.2 Entropy and Information Theory

An important aspect of the entropy concept not usually accounted for in traditional thermodynamic approaches to entropy is its information attribute.

The first to recognize the relationship between entropy and information was Szilard (18) in 1929, who related the usage of information to the production of entropy. This was approximately twenty years before the development of information theory by Shannon (19) in 1948, and its value has only recently been recognized. Shannon's contribution to science was significant, since for years investigators had tried to formulate a useful

.

measure of information for communication engineering (20). Several names stand out in the early years: Hartley (21) with his theory on information transmission, using the logarithm of number of symbols as an informational measure, and Gabor (22), working on time-frequency uncertainty and the logon concept. However, it was Shannon who clarified the confused situation with his theory.

Like Hartley, Shannon used a logarithmic measure of the number of symbols as his measure of information. Formally, Shannon's information measure for the j<sup>th</sup> symbol,  $h_j$ , is defined as the negative logarithm of the symbol's probability:

$$h_{j} = -\ln P_{j}$$
 (2.2.1)

Shannon recognized his measure determined not the quantity of information the symbol conveyed, but rather the uncertainty of information. The Shannon measure can also be applied to messages; this is accomplished by determining the expected value of all symbols in the message:

$$H = \sum_{j} -P_{j} \ln P_{j} . \qquad (2.2.2)$$

The H-function has an absolute maximum when the probabilities for all the symbols are equal (23):



$$H_{max} = -\ln P_j$$
 (2.2.3)

Using equations (2.2.2) and (2.2.3) we can explore more deeply the meaning of Shannon's information measure. The notion of uncertainty is easily deduced, for as H increases, the symbols become more equiprobable and the ability to distinguish their information content decreases, or alternatively, our uncertainty about them increases. Another way to regard uncertainty is as a measure of the number of degrees of freedom. The lower the uncertainty the fewer degrees of freedom or the greater the uncertainty the more degrees of freedom the system has. The degrees of freedom concept also denotes the idea of capacity and usually "information capacity" is what Shannon's measure of information is called. Information capacity means "message variety" to the communications engineer, a useful parameter in designing communication systems.

Given the measure of information in equation (2.2.2), it is now much easier to see the relationship of entropy, in equations (2.1.9) and (2.1.21), to information.

The progression from information theory to thermodynamics is accomplished by first relating information theory to statistical mechanics, via the partition function, and thus to the various statistical-mechanical

analogs of the laws of thermodynamics. This was first done by Jaynes (24, 25) and his conclusions have since been verified by others (26, 27, 28). The agreement between information theory and thermodynamic entropy has become such a well-accepted relationship that many current textbooks on thermodynamics and statistical mechanics rely heavily on the concept of information when presenting these subjects (29, 30, 31). Perhaps the following quote of J. von Neumann can best summarize the roles of information and thermodynamic entropy (32):

The thermodynamical methods of measuring entropy were known in the mid-nineteenth century. Already in the early work on statistical physics it was observed that entropy was closely connected with the idea of information: Boltzman found entropy proportional to the logarithm of the number of alternatives which are possible for a physical system after all the information that one possesses about the system macroscopically (that is, on the directly, humanly observable scale) has been recorded. In other words, it (entropy) is proportional to the amount of missing information.

Current investigations employing information theory methodology in the study of the thermodynamics of open systems (33) and chemical systems (34) have begun to assess the use of information-entropy in engineering disciplines other than telecommunications-related areas. Both theoretical developments and wide application of the information theory methodology came about during the 1950s (35) and brought the subject out of its infancy. The 3rd London Symposium on Information Theory is an



excellent illustration of the diversity of subjects examined using the new concepts (36). The topics for papers given ranged from studies on computers, electronics, statistics and mathematics to those on animal welfare, political theory, psychology, anthropology, economics, and anatomy, which are subjects divorced from traditional communication applications. This period also saw the rise of coding theory (37, 38), an important step increasing the utility of information theory for solving the problems of a rapidly expanding telecommunications industry.

The sixties provided less innovation than the fifties, and more time for reflection on the theory's fundamental concepts and tenets (39). Emphasis was placed on coding theory, in particular on decoding algorithms (40), and these studies have remained the basic thrust of its mathematical development (41). In addition to direct telecommunication applications, information theory began to be firmly established in several other disciplines. Psychology proved a fertile area for the application of information theory, where the individual was regarded as an information processing device (42).

The economic adaptations of the theory also found acceptance. The theory of information-entropy was used as a mathematical tool for analyzing industrial

concentration (43), the future prices of stocks (44), and as an accounting methodology (45). Utilization of this concept has become so widespread that an entropy law for general economic processes has been proposed (46). Interesting, but perhaps esoteric, is the application of an information theory observer-critic to evaluation of the philosophical arguments of Aristotle (47), and the application of the theory to detective work in criminology (48). However, an important new application of information theory was in the field of biology, where many interesting new facts were discovered about the information transcription of the genetic code onto the protein space (49, 50).

To understand the information concepts which will be employed in this work, it is necessary to discuss additional terminology and theorems pertinent to the study. The best place to begin is with the extension of Shannon's information capacity, equation (2.2.2), to n<sup>th</sup> order Markov processes or multivariate analysis. The idea behind this extension is that the possession of an information measure based on the joint probability of <u>B</u> and <u>E</u> could be used to construct the entropy of the word BE, that of <u>B</u> and <u>E</u> and <u>T</u>, the word BET, and so forth. Shannon called such information calculations n-gram entropies (51). Used in this manner the joint probability has a multivariate connotation, but this would change if



we considered the same message or word coming along a telegraph wire. Such a message is different because it is dynamic and can be thought of as a stochastic process. Because outcomes (words) in such a process would be discrete the formulation of  $\underline{H}$  in a Markovian sense is possible (19).

Let us first look at the bivariate case involving a pair of events and define the joint probability of the i<sup>th</sup> and j<sup>th</sup> elements:

$$P(ij) = P(i) P(j/i)$$
 (2.2.4)

where P(ij) is the joint probability, P(i) the probability of  $\underline{i}$  and P(j/i) the conditional probability of  $\underline{j}$ given  $\underline{i}$ . The bivariate information capacity, H<sub>2</sub>, equals the following (52):

$$H_{2} = -\sum_{i} \sum_{j} P(ij) \ln P(ij)$$
$$= \sum_{j} \sum_{i} P(j/i) \ln P(i)P(j/i) \qquad (2.2.5)$$

and if P(i) and P(j) are independent, H2 becomes

$$H_{2}^{I} = -\sum_{i} \sum_{j} P(i)P(j) \ln P(i)P(j)$$
$$= -\sum_{i} P(i) \ln P(I) - \sum_{j} P(j) \ln P(j). (2.2.6)$$

Note that by subtracting  $H_2$  from  $H_2^T$  we get a new measure, the divergence from independence or a measure of 1<sup>st</sup>-order



\_\_\_\_\_

Markov memory. Extending this to multivariate analysis or an n<sup>th</sup>-order Markov chain we have

$$H_{n} = -\sum_{i} \sum_{j} \dots \sum_{n} P(i)P(j/i) \dots P(n/N-1) \ln P(i)P(j/i)$$
$$\dots P(n/n-1)$$
(2.2.7)

and if all the probability sets are independent:

$$H_{n}^{I} = -\sum_{j}^{N} P(I) \ln P(i) - \sum_{j}^{N} P(j) \ln P(j)$$
  
- ... -  $\sum_{n}^{N} P(n) \ln P(n)$  . (2.2.8)

The difference between  $\textbf{H}_n^{\text{I}}$  and  $\textbf{H}_n$  would be the same as an n<sup>th</sup>-order measure of Markov memory.

The information-entropy measure expresses a capacity for freedom or variety and is sometimes referred to as "potential" information. Often, it is desirable to speak in terms of the order or "stored" information, I<sub>s</sub>, of a system. Logically, the order or stored information is the difference between the maximum disorder or entropy of the system and its actual entropy:

 $I_{s} = H_{max} - H$  (2.2.9)

Obviously, the notion of stored information can be extended to any of the multivariate cases, giving a measure of order for each case of dependence.



The idea of "stored information" in information theory terminology is usually conveyed by the concept of redundancy, <u>R</u>. Redundancy, as explained by Weaver (53), reflects that fraction of the message which is ordered or repetitive:

$$R = I_{s}/H_{max} = 1 - H/H_{max}$$
(2.2.10)

The error in communicating a message is reduced as the redundancy increases. Thus, the redundancy concept provides an indication of the reliability of the system.

Noise and channel capacity are two other common terms. The channel capacity,  $\underline{C}$ , is the maximum rate at which information or entropy flows through a channel (the physical medium of information transmission). Channel capacity has the units symbols per unit time (54):

$$C = ln (n)/t$$
, (2.2.11)

where ln n is sometimes referred to as the entropy of one channel. Noise is one of the limiting factors in the efficiency of transmission through a channel; it is the error between the message sent and that received. The processes of encoding and decoding the message are the pertinent factors determining the noise of a channel. One of Shannon's more important theorems describes the potential for reducing the noise of a channel (19):



Let a discrete channel have the capacity C and a discrete source the entropy per second  $\underline{\mathbb{H}}$ . If  $\underline{\mathbb{H}} \leq \underline{\mathbb{C}}$  there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If  $\underline{\mathbb{H}} \geq \underline{\mathbb{C}}$  it is possible to encode the source so that the equivocation is less than  $\underline{\mathbb{H}} - \underline{\mathbb{C}} + \varepsilon$  where  $\varepsilon$  is arbitrarily small. There is no method of encoding which gives an equivocation less than  $\mathbb{H} - \underline{\mathbb{C}}$ .

The theorem implies that we cannot eliminate noise in the channel but we can "learn to live with it."

A question generally addressed in information theory concerns the capacity of a set of symbols (words) in a code (language) to transfer information based on their respective durations (lengths). The duration or length of a word is related to its cost, because the more symbols needed to convey a bit of information, the greater the cost and the less efficient such a transfer becomes. One might suspect then that the frequency of a word in a language would be related to its cost; the longer words being less frequent and the shorter ones more frequent. An analysis by Mandelbrot (55) showed the most efficient coding scheme per unit cost satisfied the aforementioned suspicion. However, the result achieved by Mandelbrot by an information theory approach had already been realized years before by Zipf (56, 57) through empirical analyses of language. The principle or law discovered by Zipf and rationalized by Mandelbrot can be expressed either by:


- a) a relation between the frequency of occurrence of an event and the number of different events occurring with that frequency, or
- b) a relation between the frequency of occurrence of an event and its rank when the events are ordered with respect to frequency of occurrence.

Zipf's law is a power law which can be linearized by employing its logarithmic form. Mathematically, the law states that the logarithm of a word's rank in a language or code equals the negative of the logarithm of its frequency plus a constant equal to the logarithm of the frequency of the word with rank one.

Mandelbrot's method for deducing Zipf's law began by calculating the best probability rule for words of a varying cost,  $c_j$ . The idea for obtaining this rule was very similar to Boltzman's for finding the maximum of ln P or the maximizing of the number of microstates, yielding the most probable distribution. <u>P</u> was interpreted here as the total number of different messages from N words:

$$\ln P = -N \sum_{j} P_{j} \ln P_{j}$$
 (2.2.12)

and was maximized by the Lagrange's multiples method with the constraints that the sum of the P<sub>j</sub>'s equals one and that the total cost of the message of <u>N</u> words, C<sub>N</sub>, equals the sum of the costs for each j<sup>th</sup> symbol:

$$C_{N} = \sum_{j} n_{j} c_{j} = N \sum_{j} P_{j} c_{j}$$
 (2.2.13)



The result was that probability of the j<sup>th</sup> word must be related exponentially to the j<sup>th</sup> cost to get the maximum number of different words for a given cost:

$$P_{j} = \exp(-bc_{j})$$
 (2.2.14)

where b is a constant.

The next step was to find the number of words N(c<sub>j</sub>) of cost c<sub>j</sub>. This problem boiled down to a finite difference problem:

$$N(c_j) = \sum_{k} n_k N_k (c_j - c_k)$$
 (2.2.15)

which states that any one of the  $n_k$  words can be used to construct a message of cost  $(c_j - c_k)$  to build a word of total cost  $c_j$ . For a stable code, the finite difference solution of equation (2.2.15) is:

$$N(c_{j}) = A_{1} \exp (b'c_{j}) + A_{2}$$
 (2.2.16)

where  $A_1$ ,  $A_2$ , and b' are constants, and where inverse b' times the logarithm of  $A_1$  equals the negative of the cost of the initial condition,  $c_1$ .

By solving both equations (2.2.14) and (2.2.16) for  $c_j$ , assuming  $A_2$  equals zero and sorting by increasing cost or rank, the order of  $N(c_j)$  as determined by its cost is:



$$\frac{b}{b^{T}} \ln \left( \text{Rank} \left[ N(c_{j}) \right] \right) = \frac{b}{b^{T}} \ln A_{1} - \ln P_{j}$$
 (2.2.17)  
$$\frac{b}{b^{T}} \ln \left( \text{Rank} \left[ N(c_{j}) \right] \right) = -bc_{1} - \ln P_{j}$$
  
$$= \ln P_{1} - \ln P_{j}$$
 (2.2.18)

which is a generalized form of Zipf's law and dictates an ordering for the number of words of cost  $c_j$  in a message of  $\underline{N}$  words, which maximizes total message variety (efficiency).

Recently, Kozachkov (58) showed that a ratio of b/b' equal to one maximized the total message variety. He stated that the overall number of different messages is:

$$P = \sum_{j} P_{j}P = P \sum_{j} P_{j}$$
(2.2.19)

and by determining  $\textbf{P}_{j}$  from equation (2.2.18) and substituting he got:

$$P = P \sum_{j=1}^{J} P_1 N(c_j)^{-(b/b')}$$
$$= P' \sum_{j=1}^{J} N(c_j)^{-(b/b')} . \qquad (2.2.20)$$

Then he calculated the sum over J different words for three cases: b/b' greater than one, equal to one, and less than one:



$$P = \frac{P'}{1 - (b/b')} J^{1-(b/b')} , b/b' < 1$$
$$= P' \ln J , b/b' = 1$$
$$= \frac{P'}{(b/b') - 1} , b/b' > 1 . \qquad (2.2.21)$$

The maximum P clearly is for b/b' equal to one as J approaches infinity, and we can thus write Zipf's law:

$$\ln P_{j} = \ln P_{l} - \ln (Rank_{j})$$
 (2.2.22)

Kozachkov said that when a hierarchical structure or system followed Zipf's law with a slope of minus one, its organizability was maximum because the information capacity at every level in the hierarchy was maximized relative to the overall information capacity. This principle was recently used as an indicator of national citysize integration (59).

The far-reaching expressions of Zipf's law in nature are very striking phenomena. Zipf himself expanded the scope of his studies beyond that of word distribution to distribution of interval frequency in classical music, city-size frequency, product-manufacturing frequency, retail store frequency, job-occupation frequency, newspaper circulation frequency, charge account frequency in department stores, frequency of telephone messages through interchanges, and other examples which obeyed his rule (57).

Many empirical laws developed by others are Zipfian; Pareto's law of income distribution (60) being one of the more notable, has often been cited as instrumental in formulating the graduated income tax structure of today. The Lotka distribution law (61) deals with the frequency of scientific writing. The biologist's law of allometric growth, and allometry in general (62), are good examples of the organization of parts within the organism working through self-regulation for the benefit of the whole. The dose-response curves (63) demonstrate the inherent ability within an organism to interpret an incoming coded chemical message and elicit a response dependent upon the message magnitude (chemical frequency). Zipf's law is a powerful tool for assessing the organizability of various physical, biological or social systems.

The importance of information-entropy criteria in statistical inference was demonstrated by Tribus (64). He utilized the principle logically set forth by Jaynes (65) and Cox (66) that the maximum entropy formalism elicits the minimally prejudiced probability distribution. This is true since the maximum entropy state is achieved when the hypothesis favors no inference more than another. Tribus then proposed a method for calculating the probability distribution which would maximize the entropy function under the constraints imposed by the standard statistical distributions such as uniform,



exponential, gaussian, gamma, beta, etc. He then stipulated that the particular probability density distribution which maximizes the entropy function yields the minimally prejudiced probabilities. On this basis he formulated an entropy inference test:

$$\Delta S = N\left[\sum_{i} P(i) \ln P(i) + \sum_{j} P(J) \ln P(J) - \sum_{i} \sum_{j} P(ij) \ln P(ij)\right] \qquad (2.2.23)$$

which, if one recalls equations (2.2.5) and (2.2.6), is the difference between the maximum bivariate entropy,  $H_2^I$ , and the distribution bivariate entropy,  $H_2$ , quantity multiplied by N. Thus, this entropy inference test measures the independence between two distributions where independence between the two sets implies that there is no information difference between them. If one of the distributions is prejudiced, then equation (2.2.22) measures the bias in our observed distribution. Tribus showed that if the information difference between the two distributions was small, the quanity  $-\Delta S/N$  equals the Chi-square statistic.

The importance of Tribus' work and other information theory applications is that they demonstrate the role of information in our world. The individuals mentioned in this section have uncovered that role: for

example, Jaynes in statistical mechanics, Mandelbrot in Zipf's law, and Tribus in statistical inference. In subsequent chapters, I will expand the perspective on information into the area of nutrition to more fully examine the utilization of information by the living system.

### 2.3 Entropy, Information and Biology

The validity of applying information theory methodology to other fields, aside from those directly related to communication systems, was a point of some debate after Shannon's inception of the approach (67). Some felt that information theory was an approach which could only be justifiably used in telecommunication applications. However, such a viewpoint defines the theory's value only in terms of its immediate success. One cannot deny communication engineers their welldeserved credit for laying the sound mathematical foundations of the methodology, but such limited range for application unduly restricts information theory development to the parochial aspirations of this discipline. Unlike the communication engineers, physical, biological, and societal scientists could not as easily put their encoders and decoders on the table and confirm the theoretical predictions of the theory. Because the biophysicist could not take the cell's DNA and examine

its sequence nor the psychologist take a brain apart to examine its neuron network, these early ventures of information theory often resulted in frustration. This frustration led to a decline in interest that has slowly begun to be reversed with the determination of biochemical and biophysical structures and functions of living systems.

The biological field is intimately associated with the field of thermodynamics, and the important roles played by entropy, information, order and control are recognized throughout the discipline (68, 69, 70, 71). The state of knowledge in several biological fields has reached the level where application of the informationentropy concept can have and has had a significant impact on the interpretation of experimental studies.

A significant area of information theory application in biology is in the field of neurophysiology. The theoretical basis for applying information theory to the nervous system was put forth by von Neumann in the midfifties (72). His approach drew analogies between an information synthesis of a reliable system's unreliable components and neurological systems. Since this work, experimental studies have tended to support the utility of information analysis in neural systems. These applications have studied the encoding mechanism (73), transmission and multiplexing of neural information (74, 75),

and the general physiology of nervous cells (76) and systems (77).

Another prime arena of biological information theory applications is genetics, and I would like to present one of the major works on the subject as an illustration of the potential the entropic approach possesses. At the beginning of the last decade, a new understanding arose concerning the relationship between the structure of DNA and that of proteins (78). The genetic code, as this relationship is commonly called, a universal biological language for the storage and transmission of cellular information essential to metabolism and behavior, was deciphered (79). Code words are formed by a sequence of three nucleic acids which link together forming a strand of DNA. Each sequence translates into an amino acid which during transcription of the code is catalytically joined to others to form proteins.

An impressive study on the genetic code has been done by Gatlin (80). Her work examined the univariate and bivariate information capacities, equations (2.2.2), H, and (2.2.5), H<sub>2</sub>; the stored information, equation (2.2.9), I<sub>s</sub>; and redundancy, equation (2.2.10), R, of nucleic acid sequences in DNA. The results of her examination have led to new insights not only on the grammar aspect of the genetic code but also on the evolutionary process in nature.



Recall that  $I_s$ , stored information, is our entropic measure for divergence from equiprobability or equality in the univariate case or divergence from independence in the bivarate case. Let us denote  $I_{s_1}$  as univariate divergence and  $I_{s_2}$  as bivariate divergence. Both these measures have a special meaning with regard to language.  $I_{s_1}$ , the divergence from symbol equality, is the main determinant in a language of its message or word variety. The frequency of a language's symbols thus dictates the available vocabulary.  $I_{s_2}$ , the divergence from independence, is the  $1^{st}$ -order measure of a language's grammar. The degree of dependence imparts redundancy or fidelity into the message by dictating the symbols' relationships to each other (i.e., grammar).

Gatlin calculated  $I_{s_1}$  and  $I_{s_2}$  based on the percentage of guanine and cytosine, for phage, virus, bacteria, plant, insect, and vertebrate organisms. Of course, univariate information-entropy studies had been done on DNA before, but not bivariate. The union of univariate and bivariate information-entropy concepts presented in the format of a language study is a significant advance in genetics, for it gives a new methodology for interpreting the biochemistry of cellular information storage which would not be possible without information theory.



How much further this new methodology might be applied was also demonstrated by Gatlin in this study in her application of the results to the evolutionary [Such an application of information theory had process. been previously suggested (81).] She was able because of her application of information-entropy measures to hypothesize a new theory about the course of evolution in nature. The name given to this theory was Shannonian evolution. Using the measures  $I_{S_1}$  and  $I_{S_2}$ , Gatlin made several important observations. One was that the nonvertebrate species showed considerable variation in the frequency of guanine plus cytosine, whereas the vertebrates displayed little variation. Also, it was noted that the invertebrates had a significantly greater variation in both I<sub>S1</sub> and I<sub>S2</sub> than the vertebrates. After extensive analysis of the situation her conclusion was that the evolution from invertebrate to vertebrate life forms has proceeded in two phases. The first was where  $I_{s_2}$  decreased; the second, where  $I_{s_2}$  increased.

Such an evolutionary course is significant in the context of information theory because the  $I_{s_2}$  governs the relative degrees of variety and fidelity in a code. If we assume that the overall redundancy or order in the code increases consistently as one advances up the evolutionary hierarchy, then, the first phase of decreasing  $I_{s_2}$  can be regarded as the period when  $I_{s_1}$  is increasing.

 ${\rm I}_{{\rm S}1}$  affects the message variety of the code and consequently this first evolutionary phase can be looked on as a search for an optimal alphabet for message variety. The second phase of increasing  ${\rm I}_{{\rm S}_2}$  depicts an increase in the grammar or dependence of symbols in the code. This evolutionary process is similar to a child learning how to read and write. First, the alphabet is taught with simple spelling (a form of grammar). After the alphabet has been mastered (end of evolutionary phase one), the development of reading and writing skills involves learning increasingly more grammar,  ${\rm I}_{{\rm S}_2}$ , as advanced spelling, syntax, etc. (evolutionary phase two).

Gatlin's explanation of evolution through the information theory allows us to understand Darwin's theory in the context of modern evidence in genetics. Other biological information systems amenable to information-entropy analysis exist. If the genetic system is a living information storage system, the metabolic or nutrition system can be regarded as the maintenance system of an organism. Such a nutritional system operates by encoding information (food) it receives, and channeling it for use in growth. Although simply stated, the control of the various metabolic information processes is very complex and the information messages are not as neatly visualized here as in the genetic system. However, by employing an information-entropy analysis of nutritional



systems, an integrated format uniting information storage by the genes and information transmission in the metabolic control process can begin to be understood.

# 2.4 Entropy, Information and Nutrition

The aim of this thesis is to employ the concept of entropy in the context of information theory and to use this measure, information-entropy, to analyze various metabolic processes. The first question to be addressed is whether Second Law principles hold for living systems. Schröedinger (82) stated that Second Law behavior does hold for these systems, but gualified this statement by claiming life to be a steady-state process which preserves the entropy of the individual organism at the expense of increasing the entropy of its environment. Essentially, the organism maintains itself by consuming low entropy substances and transforming them into higher entropy compounds. Given that such thermodynamic behavior is valid for biosystems, can we extrapolate from what could be called an energy-entropy basis to an information-entropy basis?

A direct translation of the phenomenological laws of the thermodynamic branch to the information branch was thought by von Neumann to be consistent and a logical step. He expressed this opinion in the following manner (83):



There is reason to believe that the general degeneration laws, which hold when entropy is used as a measure of the hierarchic position of energy, have valid analogs when entropy is used as a measure of information. On this basis one may suspect the existence of connections between thermodynamics and new extensions of logics.

Fong (84) also perceives congruency between the thermodynamic and information behaviors of biologically active systems, finding the laws for the creation and dissipation of information consistent with those of Prigogine's (85) thermodynamic theory of structure, stability and fluctuations.

A biological dissipative process can best be understood in processes such as catabolism. By studying the complete catabolism of alanine in the mammalian body, it can be demonstrated through the use of an informationentropy approach that a dissipation of information occurs during catabolism as one would expect a dissipation or increase in entropy to happen. The formula for the complete respiration (catabolism) of alanine is (86):

4  $CH_3CH(NH_2)COOH + 12 O_2 \rightarrow 10 CO_2 + 10 H_2O$ 

 $+ 2 CO(NH_2)_2$  (2.4.1)

Using equation (2.2.2) to calculate <u>H</u>, our informationentropy measure, the molecular information in the above process will be dissipated if H(products) is greater than H(reactants). The molecular probability on each side of



\_\_\_\_\_

the equation can be equated, using the respective mole fractions of each compound. The following informationentropy measures can be calculated from equation (2.4.1):

and

where  $\underline{k}$  is a constant factor for converting from natural logarithms to base 2 logarithms. Results in information theory are typically expressed as the number if binary digits, termed "bits."

The amount of information present or stored in the system is given by equation (2.2.9), and  $H_{max}$  can be calculated when the five different molecular species in the above catabolic reaction are equiprobable (i.e., equal mole fractions). The informations of the respective systems are:

$$I_{s}$$
(reactants) =  $H_{max}$  - H(reactants) = 2.322 - 0.811

and



J,

The information change in going from reactants to products is  $I_s$  (reactants) minus  $I_s$  (products), which equals 0.54 bits/molecule and indicates that information is being dissipated.

Supposing that the basic dissipative laws of thermodynamics can be transferred to those of information, a detailed inspection of other information-entropy processes is necessary. Coding of information is an attribute commonly seen in biological systems. An example of a biological code is the genetic code.

However, information coding is not as readily seen in nutrition as in genetics, the reason being that the nutritional control system is a complex information hierarchy. There is not a universal code such as that in genetics which translates through a well-defined geneprotein channel, but rather, many different codes and channels make up the nutritional system. Perhaps the most obvious code of relevance to nutritional studies is that which can be termed the "protein code" (87). The term "protein code" arises from the relationship between the genetic code and proteins. If one assumes that DNA is a coded sequence of nucleic acids, it logically follows that the amino acid sequence generated by the gene itself is coded in some manner. Therefore, the protein code can be envisioned as providing the basis for the creation of chemical informational molecules whose



-----

function is dependent upon protein structure (e.g., enzymes).

Whether a protein can be formed from a nucleic acid message depends upon the availability of amino acids within the organism. Those amino acids which cannot be synthesized by the cell must come from the diet, and the proportion of such essential amino acids in the diet, as well as quantities of nonessential amino acids, will affect synthesis of cellular protein. This interrelationship provides a basis for justifying application of information-entropy to nutrition and for relating it to previous work on information theory in biology.

Of course, the protein system is but one system for metabolic transformations. However, the above discussion allows one to visualize how entropy and information concepts can be incorporated into an analysis of nutritional systems. In the following chapters, I will present more detailed discussions and analyses of the relationship of amino acid nutrition to overall protein metabolism, and of some aspects of carbohydrate biochemistry, with the aid of information theory. Through these studies I hope to elucidate the importance of the concept of information-entropy in nutritional systems.



-

#### CHAPTER III

### INFORMATION AND THE QUALITY OF PROTEINS

Nutritionists have developed many concepts over the years, both qualitative and quantitative, to express the value or worth of a food protein. These standards have been formulated by using combinations of chemical and biological analyses. This chapter will present an information-entropy approach for ascertaining a nutritive protein code and then show the relationships among the various indicators of protein quality and the informationentropy of the diet.

# 3.1 Indices of Protein Quality

Before applying the principles of information theory to protein quality, a description of the most common indices of protein quality is in order. The following concepts will be discussed: biological value, digestibility, net protein value, net protein utilization, protein efficiency ratio, chemical score, and essential amino acid index. Although it is not a complete or exhaustive list of concepts for assessing protein quality, this set of indices is representative of the more commonly used parameters.



.

.

The "biological value" (BV) is one useful estimate of protein quality, involving a nitrogen balance approach. This measure was defined in 1909 by Thomas (88) as the fraction of absorbed nitrogen retained within the organism for maintenance and growth. It may be expressed mathematically as (89):

$$BV = \frac{N_{I} - (F - F_{k}) - (U - U_{k})}{N_{I} - (F - F_{k})} , \qquad (3.1.1)$$

where  $N_I$  is the nitrogen intake, F is fecal nitrogen,  $F_k$  is endogenous fecal nitrogen, U is urinary nitrogen, and  $U_k$  is endogenous urinary nitrogen. The endogenous fecal and urinary nitrogen can be determined by finding a nitrogen-free diet or one containing a small amount of high quality protein (90). Estimates of biological value which do not correct for endogenous nitrogen losses are termed "apparent biological values."

"Digestibility" (D) is probably one of the oldest qualitative indicators used in nutritional studies. It denotes the fraction of the food nitrogen which is absorbed, and is calculated (89):

$$D = \frac{N_{I} - (F-F_{k})}{N_{I}}$$
(3.1.2)

Like biological value, digestibility is a nitrogen balance index, and is classified as "true" or "apparent"



.

depending upon the inclusion or exclusion of endogenous nitrogen losses in its determination.

Another nitrogen balance method which is a combination of the previous two indices was put forth by Bender and Miller in 1953 (91). Essentially, this new index, originally called "net protein value" (NPV), is equivalent to biological value times digestibility, and expresses the amount of nitrogen retained divided by the total nitrogen intake:

$$NPV = \frac{N_{I} - (F - F_{K}) - (U - U_{K})}{N_{I}} .$$
 (3.1.3)

Several years later, Bender and Miller proposed a shortened method for determining what is effectively the same quantity as net protein value. The difference was that this new index was approached through a carcass analysis method rather than by nitrogen balance; the name coined for this parameter was "net protein utilization" (NPU) (92). Net protein utilization was defined:

$$NPU = \frac{B - (B_k - N_{Ik})}{N_I} , \qquad (3.1.4)$$

where B is the body nitrogen of the animals fed the test protein, and  $B_k$  and  $N_{Ik}$  are the body nitrogen and nitrogen intake of the group fed the nonprotein diet.



All the aforementioned tests for protein are usually noted as being conducted either under "standardized" or under "operative" conditions (89). Standardized measurements are those made under maintenance conditions, whereas operative ones are those made under other defined conditions. Sometimes a suffix indicating the percentage of protein in the diet is used (e.g., NPU<sub>10</sub>). These are important constraints to recall when interpreting these tests, for the quality of the protein depends greatly on the purpose for which it is required (e.g., growth or maintenance).

Typically, net protein utilization and net protein value are taken as measurements of the same quantity, and are not distinguished between in the literature (93). However, for my purposes a distinction will be made. The term "net protein value" will refer to those measures calculated by multiplying digestibility times biological value (i.e., those done by a balance method), while "net protein utilization" will denote measures determind by a carcass analysis method.

The final biological estimate of protein quality to be presented here is the "protein efficiency ratio" (PER). It is a parameter proposed in 1919 by Osborne <u>et al</u>. (94), and defined as the "gain in body weight divided by weight of protein consumed." This is a very popular index, primarily because of the ease with which


it can be determined. Previously, the determination of this ratio was conducted at several levels of nitrogen intake. In this manner, an optimal level of protein intake could be identified for a maximum gain in weight. Generally, good correspondence between gain in body weight and gain in body protein exists, however PER is not always an acceptable evaluation procedure (95), and is not as reliable an indicator of protein quality as the other indices.

Although these indices are determined experimentally in guite different ways, biological value, net protein value, and net protein utilization are based upon the same criterion, retained nitrogen, and should measure essentially the same thing (96). The protein efficiency ratio would be an approximate measure of this criterion, also. Table 3.1.1 gives a listing of 21 different food proteins taken from an FAO compilation of biological data (97), for which the scores of the above four indices were found (protein level of diets from which scores were derived was 10%). The table includes both the actual score and a ranking of the proteins based upon their respective scores. Table 3.1.2 lists linear correlation coefficients (98) which were calculated using paired scores of the indices, and all regressions are significant at P < 0.01. This crosscorrelation analysis shows a very good relationship



	Biological Value		Net Protein Utilization		Net Protein Value		Protein Efficiency Ratio	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Egg, whole	93.7	1	93.5	1	90.9	1	3.92	1
Wheat, whole	64.7	14	40.3	20	58.8	11	1.53	18.5
Maize	59.4	18	51.1	17	54.5	15	1.18	20
Casein	79.7	4	72.1	4	76.8	3	2.86	5
Fish, meal	81.1	3	65.8	7	76.2	4	3.42	3
Soybean	72.8	8	61.4	8	65.9	6	2.32	7
Groundnut	54.5	20	42.7	19	47.2	19	1.65	15
Sunflower	69.6	10	58.1	9	57.0	12	2.10	13
Lentil	44.6	21	29.7	21	37.9	21	0.93	21
Rice, polished	64.0	15	57.2	10	62.7	9	2.18	11
Wheat, germ	73.6	7	67.0	5	64.9	7	2.53	6
Cottonseed	67.2	11	52.7	14	53.5	16	2.25	9
Linseed	70.8	9	55.6	11.5	59.8	10	2.11	12
Sesame	62.0	17	53.4	13	50.7	17	1.77	14
Milk, whole	84.5	2	81.6	2	81.9	2	3.09	4
Beef, muscle	74.3	6	66.9	6	73.8	5	2.30	8
Lima beans	66.5	12.5	51.5	16	47.9	18	1.53	18.5
Peas	63.7	16	46.7	18	55.8	14	1.57	16.5
Pigeon peas	57.1	19	52.1	15	44.4	20	1.57	16.5
Brewer's yeast	66.5	12.5	55.6	11.5	56.1	13	2.24	10
Fish, muscle	76.0	5	79.5	3	64.6	8	3.55	2

TABLE 3.1.1.--Listings of Biological Value, Net Protein Utilization, Net Protein Value, and Protein Efficiency Ratio Scores with their Respective Rankings (Source: FAO).



	BV	NPV	NPU
NPV	0.942	1.000	0.881
NPU	0.924	0.881	1.000
PER	0.906	0.859	0.918

TABLE 3.1.2.--Matrix of Correlation Coefficients Relating Biological Value, Net Protein Utilization, Net Protein Value, and Protein Efficiency Ratio Scores.

TABLE 3.1.3.--Matrix of Correlation Coefficients Relating Biological Value, Net Protein Utilization, Net Protein Value, and Protein Efficiency Ratio Ranks (Spearman's  $\rho_e$ ).

	BV	NPV	NPU
NPV	0.911	1.000	0.865
NPU	0.901	0.865	1.000
PER	0.893	0.839	0.928

between raw scores of the various indices. On the other hand, Table 3.1.3 has Spearman's rho ( $\rho_{\rm g}$ ) (99) correlation coefficients for the ratings (P < 0.01). Basically,  $\rho_{\rm g}$  can be regarded as a regression coefficient for two ranked variables. Spearman's index additionally tells us that not only are the scores highly correlated, but so are the relative rankings we derive from them. This



lends support to the contention that all these tests are a measure of the same variable.

The previously mentioned measures of protein quality all involve a combination of biological and chemical methods of analysis. Nutritionists have arduously sought a simple chemical procedure for determination of protein quality which would be as accurate as the experimental measure of biological value. The incentive is that biological tests are expensive and time-consuming.

One of the first attempts to minimize biological testing utilized chemical score (CS). Employing the principle of the limiting essential amino acid as a justification for their method, Mitchell and Block (100) calculated a mathematical regression between their chemical scores and the biological values of 23 different proteins.

Chemical score is represented by the minimum amino acid ratio of amino acids in a test protein to those in a standard protein; it was first advanced as:

$$CS = \frac{\min(ax_j)}{as_j}, \qquad (3.1.5)$$

where min  $(ax_j)$  is the content of the j<sup>th</sup> essential amino acid which is most limiting in a test protein and  $as_j$  is the content of the j<sup>th</sup> essential amino acid in the

standard protein (usually egg), expressed in units of milligrams of amino acid per gram protein-N or grams per 16 grams protein-N.

However, the chemical score method considers only one amino acid in the protein, so a scoring method was sought that would include more amino acids of the food protein.

A variation of the above approach, which incorporates all the essential amino acids of a protein into an index of quality, was conceived by Oser (101) and has come to be known as the Essential Amino Acid Index (EAAI). This index has been used to estimate the biological value of a food protein relative to that of a standard protein.

The EAAI is basically a determination of the geometric mean of a set of ratios. These ratios are the same as those used in the chemical score procedure (i.e., the ratio of essential amino acid concentration in an arbitrary food protein  $\underline{x}$  relative to its concentration in a standard protein). The standard protein used is egg and Oser assigns to egg the biological value of 100.

The following mathematical formula is used to calculate the index:

$$EAAI = \left(\frac{ax_1}{as_1} \times \frac{ax_2}{as_2} \times \cdots \times \frac{ax_{10}}{as_{10}}\right)^{1/10} \times 100. \quad (3.1.7)$$

Oser had two additional rules he employed in determining the EAAI: (1) the maximum value of the ratio for any essential amino acid will never exceed 1.0, and (2) the minimum ratio will never be less than 0.01.

The first of these assumptions is based on the view that any quantity of an amino acid in excess of that possessed by the standard is not needed by the organism for growth. Thus, the surplus may be disregarded. In the second assumption, the justification is that there always exist certain endogenous sources of protein (e.g., intestinal enzymes, tissue degradation) which will supply some of any essential amino acid.

The limitations of amino acid scoring methods involve the factors which influence the digestibility of the protein. For example, when the essential amino acids are not completely available for metabolism, due to malabsorption or some other factor, the tendency of any index based solely upon their content is to overestimate the real biological value. Consequently, these indices are most accurate for those proteins which are more completely digestible and lose accuracy as the protein digestibility decreases.

A closing observation concerns the accuracy of these indices in assessing protein quality. In a recent appraisal of protein quality, Bender considered it sufficient to classify proteins as poor, moderate and good,

and thought that Oser's EAAI or Mitchell's chemical score were as good indices as any to use in assessing the multiplicity of protein needs (102). Thus, the salient point seems to be that the amino acid content of the food protein may be one of the best indices of protein quality available, and most certainly is a major determining factor of the biological methods presented in this section.

## 3.2 An Information-Entropy Model of Protein Quality

The purpose of this section is to view protein-, or more specifically, amino acid-nutrition with an information-entropy lens. The study of protein nutrition is extremely complex, but the main criteria for nutritional well-being are dictated by the organism's growth and maintenance requirements. As noted previously, various indices judge protein quality by estimating that fraction of the protein which is retained for growth or maintenance, depending upon experimental constraints. Consequently, any model which addresses the problem of protein quality must consider the disparity between protein needs during growth and those of maintenance, and the ways in which such variation affects protein quality.

I wish to begin my development of an informationentropy model in a discussion of the information flow from the genetic space to the protein space. My objective

is not to extensively discuss the transcription of DNA into protein, but rather to outline the processes involved. The following sequence depicts the transcription of the information from DNA (103):

- transcription from nuclear DNA template to messenger RNA (mRNA)
- (2) mRNA to cytoplasm
- (3) attachement of 30S and 50S ribosomal RNA (rRNA) subunits (called ribosomes) to mRNA
- (4) activation of amino acid by reaction with transfer RNA (tRNA) forming aminoacyl-tRNA
- (5) aminoacyl-tRNA is directed to appropriate codon on mRNA
- (6) synthesis of peptide bond by rRNAmRNA-aminoacyl-tRNA-protein complex
- (7) termination of peptide chain by chainterminating codon.

This sequence is graphically illustrated in Figure 3.2.1.

The above relates how information coded on the DNA template passes to a protein through RNA intermediaries. As was previously mentioned in sections 2.3 and 2.4, the information present in the DNA can be defined by an information-entropy measure,  $H_n$ , based on the DNA's nucleic acid frequency and sequence. This idea can be further refined to stipulate that each protein-generating DNA template has its own individual information-entropy content. If these templates can each be assumed to be structurally unique, then their information-entropy



contents would also be unique. Given the information transfer of protein synthesis, the information-entropy level of the DNA template determines the amino acid composition of the protein. From the information theory viewpoint, such a nuclear DNA template can be regarded as an information-entropy message source, and the generation of protein structure can be accomplished by a biological communication system through which the message is sent.

Five basic components make up this information communication system: (1) a message for transmission, (2) an encoding device, (3) a channel, (4) a decoding device, and (5) a message transmitted or received. The message sent over this system must be derived from a DNA template in the genetic structure of the organism. The encoding device should consist of enzymes such as RNA polymerase, which encode the DNA message into messenger RNA. I define the messenger RNA as the channel for this system, for it carries the nucleic acid message from the nucleus to the cytoplasm, where synthesis or decoding occurs. Decoding devices for the channel are the ribosomal RNA subunits, aminoacyl-tRNA, and various protein initiating factors. The decoding process seems to be the most complex step of all, involving many phases; it could be viewed as a highly redundant process to ensure accurate decoding of the message. The message is received

by the growing peptide chain, which upon completion results in the protein-coded molecular form of the nucleic acid message. The relationships described among these biological phenomena and the information communication system are illustrated in Figure 3.2.2.

One aspect of an information system thus far ignored in the discussion of a gene-protein communication system is the notion of noise. Because of the high specificity of the encoding and decoding devices (e.g., enzymes, tRNA, etc.), virtually error-free translation from the DNA to protein occurs (103). Such noiseless transmission in the system allows the information-entropy content of the DNA template to be equivalent to the information-entropy content of the protein, for with error-free transmission in communication systems, the entropies of input and output are identical (104).

The above conservation principle between source and receiver information-entropies is very important, because we can now use it to explain the changes in protein pattern requirements during growth. A rapid rate of accretion of protein begins at birth and decreases as the animal grows older (105). This rapid protein retention results both in a higher amino acid intake requirement, and in alteration of the pattern of amino acid requirements between young organisms and adults. The higher amino acid intake requirement is easily



3.2.2.--Idealized Communication System for Transmission of Genetic Information. Figure

rationalized by observation of the increased demand for these compounds in protein synthesis. However, the alterations in the pattern requirements of amino acids during growth are not so readily explained.

During growth, a phasic development of various organs (e.g., liver, brain, skeletal muscle, etc.) occurs, and each organ's growth has its own particular amino acid pattern (105). The development of these various organ systems must be caused by the expression of a particular genetic region on the chromosomes of the organism. If the assumption, previously put forth, that each such region possesses a unique information-entropy content, is valid, then our "conservation of information-entropy" principle dictates that for the gene-protein channel the information-entropy content of the corresponding protein must also be unique. Recall that uniqueness can be defined as a particular set or pattern of symbol frequencies in information theory, meaning that each unique protein has a distinct pattern of amino acid frequencies. Thus, the differences in the pattern requirements between young and adult organisms can be understood to result from the differences in the information-entropy levels of genetic expression taking place during the early and later stages of development.

The "conservation of information-entropy" principle explains how protein metabolism and amino acid

requirements can be affected by genetic expression. Now I wish to relate the gene-protein communication systems model depicted in Figure 3.2.2 to amino acid consumption by the organism. The description of protein synthesis illustrates the importance of amino acids within the cytoplasm (called the "amino acid pool"). The presence of many amino acids in the pool results from membrane transport of the plasma amino acids into the cell (106). The primary source of plasma amino acids is the diet (107). Consequently, the decoding of the genetic message is greatly dependent upon dietary amino acids, and particularly upon the essential amino acids, because they affect the amino acid pool composition and size.

Let us momentarily review the physiological phenomena involved in transmitting dietary amino acids to the tissue cell. Most dietary amino acids are found in the polymer form. The peptide bonds linking the amino acids must first be broken to free them for absorption and utilization by the organism. This bond-breaking process is termed "hydrolysis," and begins in the stomach and is completed in the small intestine (108). The freed amino acids, coming from endogenous as well as exogenous sources, and also occasionally some small peptides, are taken through the intestinal wall by several transport systems, with each system transporting only a certain

set of amino acids. After absorption the amino acids enter the portal blood.

The first major organ the dietary amino acids encounter is the liver, which plays a central role in allocating these compounds to the other body tissues (109). Approximately 70% to 100% of the absorbed amino acids are taken up by the liver. Four possible fates await the acids absorbed here: (1) catabolism, (2) synthesis into plasma proteins, (3) release as free amino acids, and (4) storage as a part of the liver's labile amino acid reserve. The last three play important roles in supplying remaining body tissues with amino acids, although complete mechanisms for accomplishing this, particularly for the plasma proteins, are not fully understood. However, free amino acids in the plasma are transported into the cell and affect the intracellular amino acid pool. Thus, the role of the liver is that of a regulator which temporarily stores the dietary amino acids until they are required by other organs.

The overall effect of the above is that the capacity of a cell to carry on protein synthesis is directly dependent upon the ability of the diet to fulfill the anabolic requirements of the organism.

From the information perspective, the above processes can be explained in terms of a communication system. First, we have an information source, the food protein,

which contains a coded message of amino acids. The message possesses a certain information capacity determined in this case by its word frequency (or amino acid frequency). The message is then encoded (i.e., protein is digested and transported) for transmission through the communiation channel (i.e., circulatory system), then decoded (i.e., transported into cell) and directed to some final destination (i.e., cellular amino acid pool). With the inclusion of the "noise" concept in the system (i.e., those inefficiencies such as the incomplete digestion of the dietary protein or poor absorption of amino acids from the qut), an essentially complete communication system has been described for the transmission of the nutritive information in a food protein to the receptor amino acid pools in the organism. Figure 3.2.3 is a representation of this system relating the aspects of nutrition and information theory.

Out of this basic concept of a communication system, will be developed some nutritional informationentropy measures. The first question concerns the nature of our measures of information. Informational units are amino acids, of which there are approximately twenty. In the cellular pool, each of these amino acids independently maintains a particular level or concentration as a function of various metabolic outlets (106).



The combinatorial approach proposed by Kolmogorov (110), a maximum entropy formalization of Shannon's method, is very appropriate for this system. In this approach we assume that a variable,  $\underline{x}$ , containing  $\underline{N}$  elements, has an information-entropy content, H(x), equal to k ln N. Note this formulation for the information-entropy of variable  $\underline{x}$  is exactly the same as Shannon's maximum entropy expression, equation (2.2.3), where all the  $P_j$ 's are equivalent and equal to 1/N. Kolmogorov expanded this approach for a set of variables,  $x_1$ , ...,  $x_j$ , ...,  $x_n$ , each capable of taking on values,  $N_1$ , ...,  $N_j$ , ...,  $N_n$ , such that the information-entropy of this set is defined:

$$H(\mathbf{x}_{1}, \dots, \mathbf{x}_{j}, \dots, \mathbf{x}_{n}) = H(\mathbf{x}_{1}) + \dots + H(\mathbf{x}_{j}) + \dots + H(\mathbf{x}_{n})$$
$$= k \ln N_{1} + \dots + k \ln N_{j} + \dots$$
$$+ k \ln N_{n} . \qquad (3.2.1)$$

Thus, for my nutritive amino acid communication system, the variables are the twenty different amino acids present in the cellular pool, where each possesses a particular magnitude dependent upon the overall metabolic state of the cell. A similar situation holds for the dietary amino acids, but the magnitude of each of these variables is characteristic of the protein fed. Now, let's formalize this communication system into the

above combinatorial format. Starting with our informationentropy source, we stipulate that a given amino acid variable, aa, has a magnitude,  $c^{x}ax_{j}$ , for food protein  $\underline{x}$ , where  $c^{x}$  is the concentration of protein in the diet (based on molar units of protein), and  $ax_{j}$  is amino acid content of the protein (based on molar units of  $aa_{j}$  per molar unit of protein). The resultant magnitude of the amino acid variable is calculated on the basis of moles of  $aa_{j}$  input to the system. Therefore, the total information-entropy of the  $j^{th}$  variable is defined:

$$H_{x}^{t}(aa_{j}) = k \ln (c^{x}ax_{j}),$$
 (3.2.2)

and a characteristic protein information-entropy, which is based upon the typical or characteristic amino acid spectrum of the dietary protein, as:

$$H_{x}(aa_{j}) = k \ln (ax_{j})$$
 (3.2.3)

For the message source, the channel transmission rate,  $I(aa_i)$ , can be defined:

$$I(aa_{j}) = \frac{1}{\Delta t} k \ln (c^{X}ax_{j})$$
 (3.2.4)

The amino acid variable at the receiving end of the communication system has magnitude am<sub>j</sub> for the j<sup>th</sup> amino acid variable. This quantity is also mole-based. If we were to consider a value for am<sub>j</sub> based on a single cell it would be that amount of amino acid necessary to



provide for all the cellular metabolic needs. However, the total nutrition of the organism must be considered, and not only one cell. Hence, the value of am, will be that amount necessary to fulfill the metabolic requirements of all the cells in the organism. The informationentropy for the receiving end is:

$$H_{r}(aa_{j}) = k \ln am_{j}$$
 (3.2.5)

Having defined the information-entropies of the source and receiver, the next aspect of the amino acid nutritive communication system to be viewed is the notion of "channel capacity." The theorem on page 26 states that to minimize transmission errors (noise), the channel capacity must be greater than or equal to that of the source. Ability to minimize transmission error is highly desirable for any organism and nature would probably not design a system which violated conditions allowing error minimization. To minimize transmission error, the relationship which holds for the entropy of the source and the channel capacity must also hold between the entropy of encoder and decoder. That is, entropy of the decoding device must be greater than or equal to that of the encoding device. Also, the information-entropy capacity of the source cannot be greater than that of the encoder. If channel capacity is much greater than that of the decoder, then decoder entropy becomes the determinant of

low error transmission. Assuming this is the case, the decoder entropy determines the channel capacity for error-free transmission. As was previously mentioned, the decoder entropy's  $j^{th}$  variable is determined by the metabolic requirements of the receiving end of the system for that amino acid. Consequently, the  $j^{th}$  channel capacity,  $C_{mj}$ , can be taken to be the information-entropy of the respective receiver:

$$C_{mj} = \frac{1}{\Delta t} H_r(aa_j) = \frac{1}{\Delta t} k \ln am_j . \qquad (3.2.6)$$

Both the channel transmission rates and channel capacities are measures of amino acid frequencies of our system. These amino acid frequencies are very much like word frequencies in any spoken language. That amino acids are words and not symbols can be argued along genetic lines. From the genetic code we know that amino acids are coded by a combination of nucleic acid symbols, as we similarly use letters to code words. In this vein, amino acids can be regarded as words. The importance of interpreting amino acids as words is that this interpretation offers the opportunity to utilize Zipf's law and obtain a cost-frequency ranking or ordering scheme.

The main proposition of Zipf's law is that the total cost for a message of N words,  $C_{\rm N},$  equals

$$C_{N} = \sum_{j} n_{j} c_{j}$$
 (3.2.7)

For our system we sum over only one  $\underline{j}$  value because we are transmitting through a one-word channel. Therefore, equation (3.2.7) reduces to:

$$C_{N} = n_{j}c_{j}$$
 (3.2.8)

Before proceeding further, let us define more fully what is meant by "cost." The total cost,  $C_N$ , can be visualized as the total time available for the j<sup>th</sup> amino acid to do an appointed number of metabolic tasks for the best growth or maintenance performance by the organism. The number of metabolic tasks which can be performed during  $C_N$  is  $n_j$ . Associated with each  $n_j$  is an individual cost,  $c_j$ , which is the average performance time for each task in the time period  $C_N$ . Now, if there is some ideal number of tasks the j<sup>th</sup> amino acid must perform during  $C_N$  and  $n_j$  is less than the ideal, an inefficiency arises. The degree to which the system is inefficient is measured by  $c_j$ , the average task performance time.

Given the above definition, let us proceed to deduce Zipf's law as Mandelbrot did. Fortunately, combinatoric information-entropy formulation is much easier to handle than a probabilistic form. Thus, the finite difference approach is not needed to obtain the costfrequency relationship. We begin by setting  $C_N$  equal to  $\Delta t$ , the time duration of our channel. By definition, the

ideal number of metabolic tasks required of  $aa_j$  over the time  $\Delta t$  is  $am_j$ , and the number which can be supplied by food protein  $\underline{x}$  is  $c^{\mathbf{x}}ax_j$ . Substituting these values into equation (3.2.8) we get:

$$\Delta t = (am_{j})(c_{mj}) = (c^{x}ax_{j})(c_{xj}), \qquad (3.2.9)$$

which becomes:

$$\frac{c_{xj}}{c_{mj}} = \frac{a_{mj}}{c_{xax_{j}}}, \qquad (3.2.10)$$

where  $c_{xj}$  and  $c_{mj}$  are the individual metabolic performance costs of source aa<sub>j</sub> and receiver aa<sub>j</sub>, respectively. The left side of equation (3.2.10) is a ranking or ordering function of the ability to perform the ideal number of metabolic tasks relative to the number permitted by dietary limitation. The ranking is absolute if  $c_{xj}$  is restricted only to integer multiples of  $c_{mj}$ , and the readily identifiable integer sequence results. The ranking is relative if  $c_{xj}$  is any other real multiple of  $c_{mj}$ .

The above formulation is exactly equivalent to Zipf's law if we use an absolute ranking condition,  $ar_{xj} \cdot c_{mj} = c_{xj}$ . Making the above substitution, and taking the logarithms of both sides, we obtain:

$$\ln \left(\frac{(ar_{xj})(c_{mj})}{c_{mj}}\right) = \ln ar_{xj} = \ln am_{j} - \ln c^{x}ax_{j}, \quad (3.2.11)$$

where  $ar_{xj}$  is the absolute rank-order of protein <u>x</u> and and  $am_j$  is the maximum  $aa_j$  frequency of the system, and  $c^{x}ax_{j}$  is the  $aa_j$  frequency of protein <u>x</u>. Equation (3.2. 11) is an exact formulation of Zipf's law.

However, because there is not sufficient evidence to assume that  $c_{xj}$  is always some integer multiple of  $c_{mj}$ , the relative ranking form of Zipf's law will be used:

$$\ln \left(\frac{c_{xj}}{c_{mj}}\right) = \ln r_{xj} = \ln am_j - \ln c^x ax_j , \quad (3.2.12)$$

where r<sub>xi</sub> is the relative rank-order.

Terms in the above formula should look familiar, because after multiplying by a constant factor,  $k/\Delta t$ , the formula becomes the difference between channel capacity and the transmission rate. This allows one to see the continuity between the operation of the nutritive amino acid communication system and a metabolic cost-frequency ranking dictated by Zipf's law. The objective now is to utilize this concept of relative rank-order to relate our information-entropy analysis to some of the indices of protein quality discussed in the previous chapter.

We start by taking the antilogarithm of equation (3.2.12) and shifting around some terms:

$$c^{x}ax_{j} = am_{j}/r_{xj}$$
 (3.2.13)

Recall that  $c^{x}ax_{j}$  is the greatest possible quantity of the j<sup>th</sup> essential amino acid from food protein <u>x</u> that can be utilized by the organism. For the time being, let us assume that all of the j<sup>th</sup> amino acid content of protein <u>x</u> is utilized for protein synthesis by the organism, and is thereby retained. If we divide equation (3.2.13) by a factor which we assume constant over  $\Delta t$ , the total protein nitrogen intake, N<sub>I</sub>, expressions for both the biological value of the j<sup>th</sup> amino acid of <u>x</u>, BV(x<sub>j</sub>), and the net protein value, NPV(x<sub>j</sub>) (or, similarly, net protein utilization) result:

$$BV(x_{j}) = NPV(x_{j}) = \frac{c^{A}ax_{j}}{N_{I}} = \frac{am_{j}}{N_{I}} \cdot \frac{1}{r_{xj}}.$$
 (3.2.14)

This states that biological value and net protein value for a single amino acid are inversely proportional to the relative rank-orders.

By taking  $c^{x}ax_{j}$  and  $c^{s}as_{j}$  for two different food proteins, <u>x</u> and <u>s</u>, and dividing them, we get:

$$\frac{\mathbf{r}_{sj}}{\mathbf{r}_{xj}} = \frac{c^{x_{ax}}}{c^{s_{as}}} \cdot (3.2.15)$$

The following identity is seen from equation (3.2.14):

$$\frac{\mathbf{r}_{sj}}{\mathbf{r}_{xj}} = \frac{\mathrm{BV}(\mathbf{x}_{j})}{\mathrm{BV}(\mathbf{s}_{j})} = \frac{\mathrm{NPV}(\mathbf{x}_{j})}{\mathrm{NPV}(\mathbf{s}_{j})} , \qquad (3.2.16)$$



and equation (3.2.15) becomes:

$$\frac{BV(x_{j})}{BV(s_{j})} = \frac{NPV(x_{j})}{NPV(s_{j})} = \frac{c^{x_{ax_{j}}}}{c^{s_{as_{j}}}} .$$
(3.2.17)

The above equation is a combined form of Zipf's law for two variables of different rank. The important point here is that under our constraint of complete absorption and retention of the aa<sub>j</sub>, two protein quality indices have been generated from our information-entropy rule, Zipf's law.

Before removing some constraints from equation (3.2.17) and seeing what happens to BV and NPV, I wish to demonstrate an invariance of these indices when the protein concentrations of two proteins are identical. Let us assume  $c^{S}$  equal to  $c^{X}$ . The first thing which we notice is that:

$$\frac{BV(x_j)}{BV(s_j)} = \frac{NPV(x_j)}{NPV(s_j)} = \frac{ax_j}{as_j}, \qquad (3.2.18)$$

or the relationship between the protein quality indices remains unchanged. This allows us to approximate the total source entropies by the characteristic informationentropy of the food protein inputted to the system.

The first constaint I will remove is that of 100% digestibility. Removing this constraint changes equation (3.2.18) to:



$$\frac{D(\mathbf{x}_{j}) \cdot BV(\mathbf{x}_{j})}{D(\mathbf{s}_{j}) \cdot BV(\mathbf{s}_{j})} = \frac{NPV(\mathbf{x}_{j})}{NPV(\mathbf{s}_{j})} = \frac{a\mathbf{x}_{j}}{a\mathbf{s}_{j}}, \qquad (3.2.19)$$

where  $D(s_j)$  is the digestibility of protein <u>s</u> and  $D(x_j)$  is that of protein <u>x</u>. Removing the digestibility condition causes no change in our NPV relationship, but does alter our BV, leading us to conclude that NPV is an index more amenable to information-entropy analysis than BV.

From the previous equation, the chemical score index is readily derived. First, we limit consideration of amino acids to those which are essential. Then, we make protein  $\underline{s}$ , usually egg, our standard protein, against whose essential amino acid levels we will compare those of protein  $\underline{x}$ . The minimum  $ax_j/as_j$  ratio will be equal to the chemical score (CS). By taking the logarithm of equation (3.2.19) the information-entropy explanation of chemical score is made obvious:

$$k \ln [CS] = \min \left( k \ln \left( \frac{NPV(s_j)}{NPV(s_j)} \right) \right)$$
$$= \min \left( k \ln \left( \frac{ax_j}{as_j} \right) \right). \quad (3.2.20)$$

The chemical score is a measure of the amino acid channel which transmits the least information.
The essential amino acid index also can be deduced from my information-entropy analysis. Instead of searching through the essential amino acid channels for the one with the minimum unknown/standard ratio, we take the average of all the essential aa<sub>1</sub> channels:

$$= \frac{1}{10} \sum_{j \in AA} k \ln \left[ \frac{ax_j}{as_j} \right].$$
 (3.2.21)

The essential amino acid index is, thus, an average essential amino acid transmission rate through the system and if we neglect Oser's condition for rejecting that fraction of  $ax_j$  greater than  $as_j$ , the EAAI is equivalent to the average entropy over the essential amino acid set (EAA) as defined by equation (3.2.1).

Now I will summarize the material presented thus far in this section. First, a foundation for a nutritive amino acid communication system was developed, by relating a schematic of an idealized communication system to protein metabolism. Next, the type of information flowing through the system (i.e., amino acids) was discussed. The concept of "channel" was defined separately for each amino acid, the basis of this approach being the highly specific decoding mechanisms of the cell, and a



mathematical formulation of maximum and actual channel transmission was developed. The point was then made that the actual information transmission over a channel, after encoding and decoding had completely taken place for one message, reflected the amino acid frequency in the source protein message. It was then shown that the word frequency in the channel was associated with a cost in the same way that cost is reflected by word frequency in any other code. A Zipfian distribution betweeen the word frequencies of the source protein and order with respect to their costs was shown to hold. The relationship between biological value, net protein value, and rank was deduced, and resulted in general log-linear relations among word frequency, biological value and net protein value. On the assignment of various (experimentally controlled) values to the terms of our Zipfian equation, the nutritional indices of biological value, net protein value, chemical score, and essential amino acid index were obtained with the proper constraints.

This section has proposed a theoretical model for a nutritional protein communication system with analysis by information theory. The results of the analysis can be shown to correspond to some experimental and empirical protein quality indices. Such an analysis of protein nutrition is significant because it illustrates that information-entropy measures of protein nutrition can be

recognized as relevant indicators of quality. Information theory complements these chemical indices by providing them with a causal relation to nutritional protein evaluation. An analysis with experimental data of the concepts thus far presented follows.

## 3.3 Analysis of Information-Entropy Approach

An analysis of the information-entropy model will be undertaken in this section. Specifically, the model predicts correspondence between amino acid content of proteins and experimental measures of their protein quality, namely, biological value and net protein value, and this will be investigated. This analysis will be accomplished using published data on the amino acid contents of proteins and on their respective biological data.

Three information-entropy measures will be studied. Two utilize the notion of the average information-entropy of the essential amino acid set. Using equation (3.2.1) we define this variable for protein x as:

$$\bar{H}_{x}(EAA) = \frac{1}{10} \sum_{j EAA} k \ln ax_{j}.$$
 (3.3.1)

The characteristic protein form is used because the level of protein in the experimental diet was constant. In the



previous section Oser's essential amino acid index was derived from this information-entropy measure and related to a log-average of the  $aa_j$  net protein values. I shall denote the log-average of the net protein values for protein <u>x</u> as NPV<sub>x</sub>(EAA), and for protein <u>s</u> as NPV<sub>c</sub>(EAA). Equation (3.2.21) becomes:

$$NPV_x(EAA) - NPV_s(EAA) = \overline{H}_x(EAA) - \overline{H}_s(EAA)$$
. (3.3.2)

If the assumption is made that the NPV for each  $aa_j$  in the standard is equal to 1.00,  $NPV_s$  (EAA) equals zero and  $NPV_v$  (EAA) becomes:

$$NPV_{v}(EAA) = \bar{H}_{v}(EAA) - \bar{H}_{c}(EAA)$$
. (3.3.3)

The above is a logarithmic form of the EAAI, discounting Oser's rules. The antilog form of equation (3.3.3) should be an approximation of the true experimental NPV. This antilog form will be defined as  $I_x(NPV)$  if Oser's conditions are not utilized, and as  $I_x^0(NPV)$  with his conditions intact. Both are information-entropy indices.

The other information-entropy measure is one found in equation (3.2.20) which generates the chemical score index. The antilog form for the left side of that equation will be denoted as  $I_{\rm v}(\rm CS)$ .

To avoid the hazards involved with collecting values from widely scattered literature citations, I have



primarily selected data from a single extensive investigation of the amino acid content of proteins and of their effect on protein quality. The work to which I refer is that of Bjorn O. Eggum (111) in studies carried out at the Institute of Animal Science, Department of Animal Physiology, Copenhagen. However, utilizing one source also has its risks and to take into account various peculiarities or errors in Eggum's work another source on the amino acid content of proteins was employed. These data are presented in an FAO compilation of amino acid data, entitled <u>The Amino Acid Content of Foods and</u> Biological Data of Proteins (97).

Sixteen different protein diets were studied in Eggum's experiments. The essential amino acid contents of these test diets are given in Table 3.3.1 for Eggum's study, while Table 3.3.2 lists the amino acid contents for similar food proteins found in the FAO report. Unfortunately, Eggum's study did not include tryptophan values. It is readily seen by comparison that the other amino acids values in Tables 3.3.1 and 3.3.2 are very similar. Therefore, I am going to use the FAO tryptophan values with Eggum's other amino acid values in subsequent calculations of information-entropy indices. Some compromises had to be made: FAO "meat and bone meal" data was utilized in place of "meat and bone scraps"; "oatmeal" was substituted for "oats" and "dehulled oats" in the





á.

nije

Loosa

(um66a
(Source:
Proteins
of Food
gram N)
рег
(micromoles
Content
Acid
Amino
.3.1
TABLE 3

.

	әитѕұл	өпіпоілэм & Сузсеіле	эпіпоэхиТ	элііьV	əniousı	ənionəl	Рћепу1- &lanine & Тутоsine	ənibitziH	γτάτυτυς
Barley	1,577	1,366	1,889	2,843	1,753	3,383	3,136	868	1,930
Oats	1,723	1,529	1,905	2,732	1,897	3,369	3,061	068	2,178
Wheat	1,090	1,234	1,578	2,444	1,610	3,236	2,751	922	1,668
Rye	1,569	1,219	1,777	2,438	1,482	2,844	2,680	935	2,027
Maize	1,167	1,576	2,099	2,673	<b>1,</b> 796	5,051	3,141	1,060	1,553
Sorghum	783	1,046	1,894	2,886	2,145	5,542	3,357	806	1,216
Casein	3,506	1,474	2,262	3,516	2,973	5,703	3,156	1,362	1,277
Fish, meal	3,378	1,428	2,623	3,068	2,278	3,784	3,043	895	2,074
Meat and bone scraps	2,309	776	1,721	2,209	1,391	2,945	2,158	782	2,307
Soybean, meal	2,557	1,080	1,957	2,679	2,158	3,564	3,016	1,353	2,565
Groundnut, meal	1,364	762	1,438	2,091	1,663	2,940	3,299	850	3,021
Sunflower, meal	1,497	1,314	2,125	2,721	2,240	3,126	2,895	1,152	2,878
Skimmed milk powder, 50%, + dehulled oats, 50%	2,185	1,369	1,921	3,132	2,125	3,936	3,175	898	1,561
Soybean meal, 50%, + dehulled oats, 50%	1,907	1,303	1,737	2,753	1,968	3,454	2,830	1,003	2,264
Pig prestarter	2,523	1,190	2,136	3,153	2,196	4,756	3,422	931	1,600
Egg	2,843	1,867	2,697	4,023	2,745	4,241	3,783	1,024	2,207



FAO)
(Source:
Proteins
Food
of
gram N)
per
(micromoles
Content
Acid
Amino
3.3.2.
TABLE

	ənisyl	aninoidtaM & Cysteine	эпіпоэтиТ	ənileV	əuiovəlosı	əuisuəl	Phenyl- alanine s Tyrosine	өпібіјгің	9ninipaA	Tryptophan
Barley	1,480	1,286	1,739	2,691	1,711	3,178	3,015	852	1,694	471
Oatmeal	1,636	1,454	1,762	2,739	1,808	3,458	3,027	847	2,270	389
Wheat	1,279	1,295	1,536	2,356	1,555	3,179	2,741	922	1,653	333
Rye, whole meal	1,447	1,104	1,752	2,535	1,668	2,936	2,338	890	1,642	226
Maize	1,139	1,207	1,889	2,588	1,755	5,971	3,166	1,097	1,502	218
Sorghum	865	974	1,588	2,676	1,866	6,347	2,773	861	1,105	374
Casein	3,541	2,152	2,492	3,668	2,635	4,628	4,073	1,198	1,371	505
Fish, meal	3,316	1,465	2,222	2,716	2,052	3,450	2.526	1,038	2,202	294
Meat and bone meal	2,243	794	1,624	2,375	1,362	2,855	2,064	720	2,493	267
Soybean	2,728	876	2,023	2,559	2,170	3,705	2,956	1,021	2,595	392
Groundnut	1,511	808	1,367	2,231	1,612	3,050	3,231	954	4,007	312
Sunflower	1,540	1,188	1,928	2,703	2,040	3,061	2,320	937	2,869	414
Milk powder, 50% + oatmeal, 50%	2,369	1,165	1,985	3,085	1,989	4,092	3,288	1,000	1,746	414
Soybean, 50%, + oatmeal, 50%	2,182	1,390	1,893	2,649	2,158	3,582	2,992	934	2,433	391
Pig prestarter: Milk powder, 60% + oatmeal, 40%	2,515	1,377	2,030	3,154	2,228	4,219	3,340	1,030	1,641	418
Egg	2,986	2,039	2,687	3,658	2,998	4,205	3,600	982	2,186	454



50:50 mixed diets of dehulled oats with skim milk powder and soybean meal; whole meal rye for rye; soybean, groundnut, and sunflower seed for the respective meals; milk powder for skim milk powder; and a 60% milk powder + 40% oatmeal mixture for the pig prestarter.

The results of Eggum's biological tests are listed in Table 3.3.3. Two of the protein quality indices discussed in section 3.1 were measured, biological value and net protein value. Two different test animals were used: rats, each initially weighing 75 grams, and baby pigs, about 16 days old. The rats were fed 150 mg N once daily. The balance period was 5 days, and the feeding regimen was initiated 4 days before. The baby pigs were fed 6 times daily, and the protein level in their diets was 3.84% N of dry matter. The Pigs were conditioned for 6 days before the balance period, which was 4 days long.

Table 3.3.4 lists the information model's three predicted values for the experimental net protein values, which reflect various constraints upon the model (results are given on a scale of 0 to 100).  $I_x(NPV)$  is the unconstrained EAAI, whereas  $I_x^0(NPV)$  employs Oser's rules, and  $I_x(CS)$  takes account of the limiting amino acid constraint in chemical scoring. The subscripts e and f denote Eggum's and FAO data sources, respectively. A linear regression analysis (98) was done between the



TABLE 3.3.3Biological Val (Source: Eggu	ues and Net Proteir m).	<pre>Nalues of Sixteen</pre>	Test Diets of Rats	and Baby Pigs
	Biological Value (Rats)	Net Protein Value (Rats)	Biological Value (Pigs)	Net Protein Value (Pigs)
Barley	71.8	58.9	80.8	66.7
Oats	70.8	59.1	76.4	60.0
Wheat	59.0	52.9	71.2	65.2
Rye	76.7	59.0	79.7	64.5
Maize	58.1	50.9	72.6	65.5
Sorghum	52.2	44.3	73.5	62.7
Casein	71.9	72.7	84.4	83.9
Fish meal	76.3	71.6	90.1	84.0
Meat and bone scraps	48.2	42.2	64.6	55.2
Soya bean meal	62.0	56.2	71.2	64.2
Groundnut meal	60.4	55.8	59.7	54.4
Sunflower seed meal	70.7	64.9	60.4	54.6
Skimmed milk powder + dehulled oats	77.6	68.5	71.1	63.5
Soya bean meal + dehulled oats	77.4	66.5	75.1	64.7
Prestarter for baby pigs	85.2	76.7	78.1	72.2
Egg	98.6	99 <b>.</b> 5	1	8



Proteins.
Food
Different
Sixteen
for
Indices
Information-Entropy
.3.4
'n
TABLE

		Eggum	Data			FAO Da	ta	
	$\bar{H}_{x}(EAA_{e})$	I <sub>X</sub> (NPV <sub>e</sub> )	I <sup>0</sup> (NPV <sub>e</sub> )	I <sub>x</sub> (cs <sub>e</sub> )	$\bar{\mathrm{H}}_{\mathrm{x}}^{}(\mathrm{EAA}_{\mathrm{f}})$	I <sub>x</sub> (NPV <sub>f</sub> )	I <sup>0</sup> (NPV <sub>f</sub> )	I <sub>x</sub> (cs <sub>f</sub> )
Barley	10.712	76.4	76.1	55.5	10.629	72.0	71.8	49.6
Oats	10.732	77.4	77.4	60.6	10.700	75.6	75.4	54.8
Wheat	10.491	65.5	65.5	38.3	10.502	65.9	65.9	42.8
Rye	10.498	65.8	65.8	49.8	10.440	63.2	63.2	48.5
Maize	10.638	72.6	71.1	41.1	10.598	70.5	67.3	38.1
Sorghum	10.569	69.2	67.4	27.5	10.505	66.1	63.4	29.0
Casein	11.058	97.1	88.0	57.9	11.114	1001	97.9	62.7
Fish, meal	10.878	85.7	84.2	64.8	10.812	81.7	81.2	64.8
Meat and bone scraps	10.452	63.8	63.5	41.6	10.438	63.1	62.2	38.9
Soybean, meal	10.849	84.0	80.4	57.9	10.791	80.6	78.9	43.0
Groundnut, meal	10.500	65.9	63.9	40.8	10.580	69.6	65.5	39.7
Sunflower, meal	10.796	80.9	9.77	52.7	10.692	75.2	73.2	51.6
Skimmed milk powder and dehulled oats	10.777	79.9	79.9	71.2	10.801	81.1	80.9	57.1
Soybean meal and dehulled oats	10.732	77.4	77.3	64.4	10.795	80.8	79.8	72.0
Pig prestarter	10.847	83.9	82.9	63.7	10.860	84.5	84.1	67.5
Egg	101.11	100.0	100.0	100.0	11.103	100.0	100.0	100.0



5.

experimental biological values and net protein values of both rats and pigs and the three information-entropy indices. Table 3.3.5 lists the correlation coefficients (98) of this regression analysis. All correlations between the information-entropy indices and biological evaluations are highly significant (P < 0.005 for all regressions). Because the digestibility of the protein affects the model's ability to predict this parameter the correlation of the information-entropy indices is somewhat poorer for biological value than for net protein value. This was anticipated, however, and a reasonable correlation still exists between the model's predictions

TABLE 3.3.5.--Matrix of Correlation Coefficients for Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on amino acid content of dietary protein).

	Net Protein Value (Rats)	Biological Value (Rats)	Net Protein Value (Pigs)	Biological Value (Rats)
I (NPV e)	0.843	0.706	0.795	0.673
I <sup>0</sup> (NPV <sub>e</sub> )	0.898	0.794	0.817	0.729
I_(CS_e)	0.914	0.896	0.702	0.675
I (NPV )	0.856	0.704	0.786	0.634
$I_{\mathbf{x}}^{0}(NPV_{\mathbf{f}})$	0.880	0.748	0.806	0.677
I <sub>x</sub> (CS <sub>f</sub> )	0.951	0.916	0.777	0.729



and biological value. The two information indices,  $I_x(NPV)$  and  $I_x^0(NPV)$ , in which were considered the total essential amino acid contents of the proteins, gave a more consistent interspecies correlation than did the chemical scoring estimate. This seems to suggest that a total entropy criterion based on all essential amino acid channels gives better results than relying on the entropy of a single channel.

A ranking of the sixteen proteins based on their respective scores was done. Spearman's rank correlation coefficients (99) were then computed. The correlations

TABLE 3.3.6.--Matrix of Spearman's Rank Correlation Coefficients for Ranks of Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on amino acid content of dietary protein).

	Net Protein Value (Rats)	Biological Value (Rats)	Net Protein Value (Pigs)	Biological Value (Pigs)
I (NPV e)	0.824	0.629	0.546	0.513
$I_{\mathbf{x}}^{0}$ (NPV <sub>e</sub> )	0.859	0.685	0.615	0.582
ı <sub>x</sub> (cs <sub>e</sub> )	0.854	0.848	0.476	0.499
I (NPV )	0.888	0.727	0.594	0.526
$\mathbf{I}_{\mathbf{x}}^{0}(\mathbf{NPV}_{\mathbf{f}})$	0.897	0.747	0.641	0.538
I <sub>x</sub> (CS <sub>f</sub> )	0.950	0.906	0.553	0.582



among the rank-orderings, Table 3.3.6, as dictated by biological testing and information-entropy, were significant (P < 0.05) for all regressions). This analysis shows that not only does the information-entropy model generate significantly correlated scoring, but the rankings of these scores are also consistent.

Eggum determined the individual amino acid availabilities for the food proteins, and thus performance of the model was tested using quantities of available amino acids as information sources. Each of the informationentropy indices was recalculated using the fraction of the protein amino acids which was available. A linear regression analysis and rank correlation were again done, and the resultant correlation coefficients are given in Tables 3.3.7 and 3.3.8, respectively. The results of this analysis indicate that the model predictions are relatively uninfluenced by the use of the original amino acid content of the protein as opposed to the use of their respective availabilities. Only the  $I_x(CS)$  index, based on the chemical scoring assumption, exhibits a consistent improvement.

The final analysis undertaken in this section was an examination of the use of Zipf's law in the model. If a Zipfian relationship is present, a <u>log-log</u> plot of our measure of information frequency versus the ranking



TABLE 3.3.7.--Matrix of Correlation Coefficients for Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on available amino acid content of dietary protein).

	I <sub>x</sub> (NPV)	$I_{\mathbf{x}}^{0}(NPV)$	I <sub>x</sub> (CS)
Net Protein Value (Rats	.) 0.832	0.887	0.944
Biological Value (Rats	0.643	0.700	0.866
Net Protein Value (Pigs	0.811	0.864	0.805
Biological Value (Pigs	0.675	0.817	0.705

TABLE 3.3.8.--Matrix of Spearman's Rank Correlation Coefficients for Ranks of Information-Entropy Measures Versus Net Protein Values and Biological Values of Rats and Baby Pigs (based on available amino acid content of dietary protein).

	I <sub>x</sub> (NPV)	$I_{\mathbf{x}}^{0}$ (NPV)	I <sub>x</sub> (CS)
Net Protein Value (Rats)	0.782	0.812	0.918
Biological Value (Rats)	0.562	0.591	0.824
Net Protein Value (Pigs)	0.629	0.738	0.571
Biological Value (Pigs)	0.551	0.700	0.589

function should be linear. The information frequencies for our information variables are given as follows: for  $I_{v}(NPV)$  the logarithm of the information frequency is  $\bar{H}_{x}$  (EAA), while for  $I_{x}^{0}$  (NPV) the log frequency is  $\bar{H}_{x}$  (EAA), which, modified by Oser's rules, will be denoted  $\bar{H}^0_{\mathbf{x}}$  (EAA), and that of  $I_x(CS)$  is min  $H_x(aa_j)$ . The ranking function is the inverse of the experimental net protein value or biological value index (scaled to 1.0), as stipulated in section 3.2. The amino acid frequency for the chemical scoring method was calculated by multiplying  $I_x(CS)$  by the log-average amino acid frequency for egg protein, anti-log  $\bar{H}_{v}$  (NPV) of egg. The logarithmic form of the resultant frequency is denoted min  $H_x^s(aa_i)$ . This standardization procedure is done to remove the variation which would occur in the analysis if raw min  $H_{y}(CS)$  were used. The results of a correlation and regression analysis are presented in Tables 3.3.9 and 3.3.10. Table 3.3.9 gives the correlation coefficients. All correlations are at least significant at the P < 0.01 level and these results tend to support the cost-frequency behavior of Zipf's law.

However, the most interesting aspect of the analysis is found in Table 3.3.10, which lists the slopes of the Zipfian regression analyses. The information-entropy model would predict a slope of -1.00, but except for several values, our regression analyses yield slopes of approximately -0.50. Figure 3.3.1 illustrates this result

P P	1
Analvsis	The Trees of the t
(log-log)	
or Zipfian	
lation Coefficients fo	ropy Model.
E 3.3.9Matrix of Corre.	Information-Ent <sub>1</sub>
TABL.	

Net Protein Biological Value (Pigs) Value (Pigs)	-0.762 -0.634	-0.781 -0.688	-0.604 -0.581	-0.752 -0.590	-0.775 -0.639	-0.697 -0.651
Biological Value (Rats)	-0.703	-0.777	-0.876	-0.700	-0.748	-0.911
 Net Protein Value (Rats)	-0.834	-0.875	-0.884	-0.852	-0.878	-0.937
Infor-Ranking mation Frequency tion	H <sub>x</sub> (EAA <sub>e</sub> )	$\bar{H}_{x}^{0}(EAA_{e})$	min <sub>e</sub> H <sub>X</sub> (aa <sub>j</sub> )	H <sub>x</sub> (EAA <sub>f</sub> )	$\bar{H}_{x}^{0}$ (EAR <sub>f</sub> )	min <sub>f</sub> <sup>H</sup> x (aa <sub>j</sub> )

••• ··••

TABLE 3.3.10Matri Model	ix of Slopes of Zir L.	fian (log-log)	Analysis of Infor	rmation-Entropy
Infor- Ranking mation Func- Frequency tion	Net Protein Value (Rats)	Biological Value (Rats)	Net Protein Value (Pigs)	Biological Value (Pigs)
<mark>н</mark> х (еад <sub>е</sub> )	-0.532	-0.517	-0.622	-0.640
$\overline{H}_{x}^{0}(EAA_{e})$	-0.526	-0.538	-0.601	-0.654
min <sub>e</sub> H <sub>x</sub> (aa <sub>j</sub> )	-1.253	-1.429	-1.097	-1.303
$\bar{H}_{x}(EAA_{f})$	-0.573	-0.542	-0.648	-0.628
$\overline{H}_{x}^{0}$ (EAA <sub>f</sub> )	-0.608	-0.596	-0.688	-0.700
min <sub>f</sub> Hx(aa <sub>j</sub> )	-1.328	-1.487	-1.267	-1.460





Figure 3.3.1.-- $\overline{H}_{x}^{0}$ (EAA), the Average Information-Entropy, Versus Zipfian-Rank-Ordering for Net Protein Value for Rats.





for the information-entropy model of the net protein value rank-ordering of rats. This apparent error in the model is due to a constraint we still have operating, namely, total retention of all amino acids fed into the system. This effect will be explored more fully in the discussion when the model is assessed in light of all evidence.

In general, the information-entropy model correlates with both scores and rankings of such indices as biological value and net protein value. These results tend to indicate that information theory could provide a causal interpretation for underlying biological phenomena and serve as an aid in rationalizing observed nutritional behavior.


### CHAPTER IV

## INFORMATION AND THE HYDROLYSIS OF CARBOHYDRATE POLYMERS

The release of simple sugars from complex glucose polymers is directly related to the nutritonal values of these substances. A relationship between chain length and hydrolysis can be deduced by an information-entropy analysis. The organization of carbohydrate information is related to the polymer's length, and this structural information affects the rate of hydrolysis. The following presents an information-entropy approach for discerning the above relationship and verifies the analysis with experimental data on such degradative processes.

### 4.1 Aspects of Carbohydrate Structure, Hydrolysis and Metabolism

Unlike protein structure, carbohydrate structure is usually based on the frequency of only one type of chemical information, namely, glucose (112). Glucose is a hexose or six-carbon sugar. Glucose is commonly found in polymer forms of which there are two main linear classes, amyloses and celluloses. The difference between amylose and cellulose is structural: the way the glucose



monomers are bonded into chains differs. In carbohydrates the general picture is of a single information unit linked together in different ways, whereas in protein nutritions there are many different information units linked together in one way (i.e., peptide bond).

An amylose bond originates at the  $\alpha$ -position on the asymmetric carbon of glucose (113) and the cellulose at the  $\beta$ -position (114). Primarily, the linkages go from the C-1 to the carbon at the C-4 position in the adjoining molecule when the linear polymers are formed. Aside from these linear bonds branching ones also exist; the most common being the  $\alpha$ -1,6. This discussion will be limited to the linear 1-4 linkages.

Both amylose and cellulose must be degraded, so their glucose can be made available in monomer form, before these substances possess nutritional value. Degradation is accomplished by bond-specific enzymes. Those glucan hydrolase enzymes which react with the  $\alpha$ -1,4 linked glucose of amylose are known as amylases (115), while those which react with the  $\beta$ -1,4 linkages of cellulose are called cellulases (116). These glucan hydrolases operate in two different ways. The exo-enzyme mechanism attacks the nonreducing end of the polymer, cleaving off disaccharides in an endwise fashion. The endo-enzyme mechanism attacks the internal linkages of the polymer, randomly breaking it down, initially into a



mixture of di- and tri-saccharides, and finally into a mixture of mono- and di-saccharides.

The residual di-saccharides and the few trisaccharides produced by polymer degradation are readily hydrolyzed into glucose by an enzymatic class known as the glucosidases. These enzymes are specific for the  $\alpha$ - or  $\beta$ -bonds of two or three unit glucose chains and react poorly or not at all with polymers of greater chain length. Maltase is the common name of the glucosidase degrading the  $\alpha$ -linked di-saccharide, maltose, whereas cellobiase is the enzyme acting on the  $\beta$ -dimer of glucose, cellobiose.

Once glucose is obtained from the digestion of carbohydrate polymers, one of its metabolic functions is to provide the organism with energy. This energy is obtained by the breakdown of glucose into carbon dioxide and water. Two metabolic pathways are needed to derive the nutritional energy from glucose (117). The Embden-Meyerhopf pathway takes glucose and converts it into two molecules of pyruvate with some generation of biochemical energy (ATP). The pyruvate is then oxidized, with the loss of a carbon, into an acetyl group which enters the tricarboxylic acid cycle and is completely oxidized into carbohydrate and water with significant generation of ATP.

Given the above metabolic role of carbohydrates, how is their nutritional value measured? Digestibility

is the main criterion for ascertaining the value of carbohydrate. In the routine analysis of feeds, the nutritional benefit of carbohydrate results from the digestibility of two fractions, the crude fiber fraction and the nitrogen-free extract. The chemical procedure for this fractionation of feeds was developed over 100 years ago and is known as the Weende method (118). The crude fiber fraction consists basically of cellulose, lignin, and other structural polysaccharides. The nitrogen-free extract consists of amylose, sugars, lignin, and material known as hemicellulose. Generally, then, the digestibilities of cellulose and anylose components of the feed are studied. Once the digestibilities of crude fiber and nitrogen-free extract fractions are known, the contributions to digestible and metabolizable energy of the cellulosidic and amylosidic components can be determined (119). The conversion factor for carbohydrates used in calculating the digestible energy from the digestible crude fiber and nitrogen-free extract fractions is 4 kcal/g. Also, if only the crude fiber and nitrogenfree extract fractions are considered, the digestible energy equals the metabolizable energy.

A chemical analysis of the feedstuff ingested by an animal will give the necessary data on the quantity of carbohydrate in the diet. The energy contribution of these carbohydrates to the organism's metabolism is



ascertained by experimentally determining the digestibility of this fraction. Given the polymeric nature of carbohydrates and the role these polymers have in nutrition, their structure-function behavior could be patterned after an information-entropic behavior similar to that in the previous analysis of proteins. Just what informationentropic rules are followed in carbohydrate metabolism will be examined in the following section by analyzing how the polymeric structure of carbohydrates affects the rate and extent of their hydrolysis.

#### 4.2 An Encoding Model for Carbohydrate Information

In Chapter III the information-entropy approach was used to analyze the transmission of nutritional protein information. The capacity of the decoder was important in the previous analysis because it reflected the optimal metabolic requirement for amino acids or protein information. This demand had to be uniquely satisfied for each essential word (amino acid), because each was missing information required for development. The information requirements in carbohydrate nutrition differ from those for protein nutrition.

First, a carbohydrate requirement does not exist <u>per se</u>, but rather, an energy requirement does. The primary function of carbohydrates is to satisfy the energy requirement, a requirement also fulfilled by proteins and

lipids. Therefore, carbohydrates do not supply essential information as proteins do. The channel and decoding capacities are not as relevant information-entropy parameters for carbohydrates as they are for proteins. This is because the excessive carbohydrate storage capability of the organism places no limit upon the transmission of the carbohydrate message once it enters the channel. Consequently, the carbohydrate information encoded (i.e., transported) into the organism identifies the carbohydrate nutritional contribution.

In the previous section the main carbohydrate messages, amylose and cellulose, and their basic information unit, glucose, were discussed. To be properly encoded, a carbohydrate polymer message must be reduced to the monomer form. This encoding process is analogous to the enzymatic digestion of the polymer and we can think of the amylase and cellulase enzymes as encoding devices. Typically, the efficiency of the encoding process is related to the length of the message to be encoded. The longer the message, the greater the cost of encoding and the lower the efficiency.

Message length in carbohydrate chemistry is synonymous with the degree of polymerization, DP. Consider a carbohydrate source which is monodisperse (i.e., all molecules have the same degree of polymerization) (120). If N<sub>a</sub> equals the total number of glucose units in



this source, the frequency of messages,  $m(DP_j)$  with length  $DP_j$ , equals  $N_g$  divided by  $DP_j$ . Given the relationship between message length and encoding efficiency, Zipf's law will order message length with respect to encoding cost. Equation (2.2.22) gives this rank-frequency

relation:

 $k \ln m (DP_{i}) = k \ln N_{a} - k \ln (Rank_{i})$ , (4.2.1)

or, alternatively,

$$k \ln (N_{\alpha}/DP_{j}) = k \ln N_{\alpha} - k \ln (Rank_{j}) . \qquad (4.2.2)$$

Solving equation (4.2.2) for  $\operatorname{Rank}_j$  yields the relationship:

Equation (4.2.3) gives both the logarithm of the rank and the absolute redundancy, which, if divided by k  $\ln N_g$ , becomes the relative redundancy. The encoding efficiency,  $EF_j$ , equals one minus the relative redundancy, and for carbohydrates has the following form:

$$EF_{j} = 1 - \frac{k \ln DP_{j}}{k \ln N_{g}} = \frac{k \ln (N_{g}/DP_{j})}{k \ln N_{g}}$$
$$= 1 - \frac{k \ln (Rank_{j})}{k \ln N_{g}} . \qquad (4.2.4)$$



The above equation is identical to Reza's definition of encoding efficiency (121): the entropy of the original message ensemble, k ln  $(N_g/DP_j)$ , divided by the maximum information, k ln  $N_g$ , times the average length of the encoded message (equal to one for monomers). The interpretation of the encoding process for a carbohydrate message, then, is that as the cost of encoding increases (i.e., as the ranking of messages with respect to their degree of polymerization increases) the efficiency of the encoding process decreases.

Now comes the question of relating the above result from our information-entropy analysis to a relevant nutritional index. Let us begin by defining the cost associated with encoding further. Equation (2.2.13) gives the total message cost which can accommodate an information-entropy analysis. By considering a monodisperse carbohydrate message where the frequency,  $n_j$ , of words can be determined by dividing the total number of symbols (monomers) present,  $N_g$ , by the message length or degree of polymerization,  $DP_j$ , the total cost,  $C_J$ , becomes:

$$C_{j} = n_{j}c_{j} = \frac{N_{g}}{DP_{j}}c_{j}$$
 (4.2.5)

An expression for the degree of polymerization based upon the above variables is:



$$DP_{j} = \frac{N_{g}}{C_{j}} c_{j}$$
 (4.2.6)

The variable  $c_j$  is, in the context of our definition, the cost (i.e., time) per word. The total activity of an enzyme is defined as the substrate consumed per unit time (122). Consideration of the dimensions of these two variables indicates that they are inversely related;  $c_j$  is equal to the inverse of the total enzymatic activity. It is also logical that the enzymatic activity is a determinant of the cost of encoding because the encoding device for the carbohydrate message is an enzyme. Given that this relationship is correct, a link between a physical measure of carbohydrate hydrolysis (i.e., enzyme activity) and information encoding can be established. Substituting the total enzymatic activity,  $a_j$ , for a carbohydrate polymer of degree of polymerization j, we obtain:

$$DP_{j} = \frac{^{N}g}{C_{J}} \frac{1}{a_{j}}.$$
 (4.2.7)

Let us now consider two monodisperse systems of different degrees of polymerization, each with a total of N<sub>g</sub> monomers. The total costs,  $C_J$  and  $C_I$ , of each system are equal, but the individual message cost,  $c_j$  and  $c_i$ , will be different. This is logical because the message



frequency of the system possessing the lower degree of polymerization increases proportionally as its individual message cost decreases. Now, by taking the logarithm of equation (4.2.7) for the j<sup>th</sup> and i<sup>th</sup> polymer systems and calculating the difference between them, the following is seen to be true:

 $k \ln DP_{i} - k \ln DP_{i} = k \ln (a_{i}/a_{i})$  (4.2.8)

Note that a similar result is obtained by computing the difference between the  $j^{th}$  and  $i^{th}$  systems using equation (4.2.3), and results would be identical if the inverse total enzymatic activity were equal to the rank.

Assuming that the enzymatic activity does have such a relationship to rank, what is the significance? Firstly, equations (4.2.7) and (4.2.3) imply that as rank of DP<sub>j</sub> increases, the total enzymatic activity will decrease, or alternatively, the respective polymer cost will increase. Then, if we assume DP<sub>i</sub> to be greater than DP<sub>j</sub>, the ratio  $a_i/a_j$  measures the relative rate of hydrolysis as the more highly polymerized i<sup>th</sup> system is degraded to a less polymerized state, the j<sup>th</sup> system. This activity ratio is also known as the yield or recovery (122) of the activity at the n<sup>th</sup> step of a reaction, compared with some reference level. Since as the polymer is degraded, it attains a lower degree of polymerization,



the activity increases proportionally; the activity recovered by the polymer inbeing degraded can be thought of as proportional to the degree to which it has been hydrolyzed. Therefore, the activity ratio  $a_i/a_j$ , measuring the proportion of the activity recovered by the molecule during degradation, can be viewed as an estimate of a hydrolysis coefficient,  $D_i$ . Setting  $a_i/a_j$  equal to  $D_i$ , the degree of hydrolysis the i<sup>th</sup> polymer system has undergone when it possesses a degree of polymerization of j, and substituting into equation (4.2.8), yields:

 $k \ln DP_{i} = k \ln DP_{i} + k \ln D_{i}$ , (4.2.9)

which relates the information-entropy measure, message length, to the degree of hydrolysis. The hydrolytic measure of equation (4.2.9) is identical to the ratio of the encoding efficiency of the j<sup>th</sup> system to that of the i<sup>th</sup> system. Using equation (4.2.4) to determine  $\text{EF}_{i}$  and  $\text{EF}_{j}$ , the ratio equals:

$$\frac{EF_{i}}{EF_{j}} = \frac{k \ln (N_{g}/DP_{i})}{k \ln N_{g}} - \frac{k \ln (N_{g}/DP_{i})}{k \ln N_{g}}$$
$$= \frac{k \ln (DP_{j}/DP_{i})}{k \ln N_{g}} , \qquad (4.2.10)$$



which is proportional to the extent of hydrolysis,  $D_{j}$ , in equation (4.2.9).

Thus, the cost-efficiency reasoning encompassed by Zipf's law has led to ordering of the hydrolyses of carbohydrate messages based on their lengths. This relationship in turn has been shown to be identical to the encoding efficiency of the carbohydrate message, which is probably the most sensitive variable in the transmission of carbohydrate information. The agreement between this approach and experimental data will be demonstrated in the following section.

Before assessing the information-entropy method for estimating the hydrolysis of amylose, a modification is necessary because when activities of enzymes are determined, the experimental conditions are constrained so the concentration or word frequency,  $n_j$ , instead of  $N_g$ , the total monomer concentration, is constant. Thus, equation (4.2.7) becomes:

$$DP_{j} = \frac{N_{gj}}{C_{J}} \frac{1}{a_{j}}, \qquad (4.2.11)$$

where  $N_{gj}$  is the number of glucose monomers in the j<sup>th</sup> system, equal to  $DP_j$  times  $n_j$ . Given equation (4.2.11), the difference between the j<sup>th</sup> and i<sup>th</sup> polymer systems is:



k ln 
$$DP_{j} - k \ln DP_{i} = k \ln \left(\frac{N_{gj}}{C_{J}} \cdot \frac{C_{I}}{N_{gi}}\right)$$

+ k ln 
$$(a_i/a_j)$$
 . (4.2.12)

In order to see how equation (4.2.12) differs from equation (4.2.8), both the relationships between  $N_{gj}$  and  $N_{gi}$  and those between  $C_J$  and  $C_I$  must be known.

If we denote  $DP_i$  as greater than  $DP_j$ , the experimental constraint of  $n_j$  being equal to  $n_i$  implies that  $N_{gi}$  equals  $N_{gj}$  times the ratio of  $DP_i$  to  $DP_j$ . That the ratio of  $C_I$  to  $C_J$  is equal to that of  $DP_i$  to  $DP_j$  can be shown by first changing  $N_{gi}$  to  $N_{gj}$  times the  $DP_i/DP_j$  ratio and substituting it into equation (4.2.11), which allows a solution of  $DP_j$  in terms of  $C_I$ . Alternatively, equation (4.2.7) gives a solution of  $DP_j$  in terms of  $C_J$ . Equating these two expressions for  $DP_i$  shows the ratio of  $C_I/C_J$  equal to  $a_j/a_i$ , and the activity ratio given by equation (4.2.8) equal to the ratio  $DP_i/DP_j$ . Putting together these relationships, the ratio term of equation (4.2.12) becomes:

$$\begin{pmatrix} N_{gj} \\ \overline{N}_{gi} \\ \cdot \\ C_J \end{pmatrix} = \begin{pmatrix} DP_j \\ \overline{DP_i} \\ \cdot \\ N_{gj} \\ \cdot \\ N_{gj} \\ \cdot \\ DP_j \end{pmatrix} = 1 , \qquad (4.2.13)$$

the logarithm of which equals zero. Therefore, the experimental modification of making the initial



÷.

÷

und mit

concentrations (i.e., word frequencies) equal necessitates no adjustment of the information-entropy relationship and equation (4.2.9) is valid.

The remaining question concerns the validity of this information-entropy approach in determining an experimental value such as enzymatic activity. If the mathematical relations previously developed can be confirmed by experimental means, then more faith can be placed in the viability of the information approach as a methodology for understanding enzyme hydrolysis. The following section will assess the agreement of experimental data with information-entropy theory.

# 4.3 Assessment of the Carbohydrate Information-Entropy Analysis

Because the conditions of the informationentropic analysis were based upon a very selective type of experimental situation, the first part of this assessment will focus on monodisperse systems. However, the approach will later be modified so that hydrolyses of polydisperse systems can be calculated. The total activity of the enzyme will be determined from velocity of the reaction (moles/unit time) as given by the Michaelis-Menten equation (123), and using the kinetic constants characteristic of the enzyme under study.

The data for amyloses were taken from a paper by Husemann and Pfannemuller (124), who experimentally



determined the kinetic constants  $V_m$ , maximum reaction velocity, and  $K_m$ , the Michaelis-Menten constants for two amylases,  $\beta$ -amylase and phosphorylase synthetase (sources of the enzymes in the studies were not discernible). Both are exo-enzymes:  $\beta$ -amylase cleaves off maltose units from the non-reducing end of the polymer, while phosphorylase cleaves or adds glucose-l-phosphate units to the ends of polymers (125). The experiment was done with amylose having degrees of polymerization from 750 to 3,815 where the molecular weight distribution for each polymer was narrow (i.e., approximating a monodisperse solution). These amylose chains served as sites of degradation for  $\beta$ -amylase action or sites of synthesis for phosphorylase. The substrate concentration used in calculating the total activities of the enzymes was 0.06 M. Table 4.3.1 (page 111) presents the experimental kinetic data for these enzymes while a graphic display of these relationships to DP, for  $\beta\text{-amylase}$  can be found in Figure 4.3.1 (page 109). The results of a correlation and regression analysis (P < 0.05) between total enzymatic activity and degree of polymerization are found in Table 4.3.3 (page 112).

Data on monodisperse solutions of carboxymethyl cellulose for four different cellulase complexes, after an experiment by Almon and Eriksson (126), were used to investigate the chain length-activity relationship for





Figure 4.3.1.--The Degree of Polymerization Versus Enzymatic Activity for  $\beta$ -Amylase.





Figure 4.3.2.--The Degree of Polymerization Versus Enzymatic Activity for Cellulase <u>A</u> (<u>Penicillium Notatum</u>).



DP j	β-Amylase			Phosphorylase		
	к_*	v_**	a,**	ĸ <sub>m</sub> <sup>†</sup>	v <sub>m</sub> <sup>++</sup>	a <sup>††</sup>
750	66.6	10.3	10.29	19.0	22.2	22.19
1,775	39.4	6.25	6.25	13.0	14.7	14.70
2,875	29.6	4.0	3.99	6.75	8.04	8.04
3,815	23.3	3.45	2.92	3.25	4.50	4.50

TABLE 4.3.1.--Degree of Polymerization and Enzyme Kinetic Data of Amylose.

\*micromoles maltose.
\*\*micromoles maltose/sec.
†micromoles glucose.
††micromoles glucose/sec.

TABLE 4.3.2.--Degree of Polymerization and Activity Data of Cellulose, with activity in (moles/sec.)  $\times 10^{-9}$ .

DP j	Cellulase A*	Cellulase B**	Cellulase C <sup>†</sup>	Cellulase D <sup>††</sup>
112	124	109	126	90
118	85	68	66	68
128	158	185	248	156
211	58	60	80	86
291	69	57	57	60
323	42	25	35	30
351	37	16	31	29
625	17	10	19	12
871	6.2	6.2	7.3	7.1
885	8.0	5.9	5.9	5.5
988	14.3	6.5	8.9	11.2

\*Cellulase purified from Penicillium Chrysogenim Notatum. \*\*Cellulase dialyzed from Aspergillus Oryzae Niger. <sup>†</sup>Cellulase partially purified from Aspergillus Oryzae Niger. <sup>†</sup>Cellulase purified from Stereum Sanguinolentum.



Enzyme	Cor: Coe:	relation Eficient	Slope	x-intercept
β-Amylase		0.99	-1.33	14.41
Phosphorylase		0.95	-0.97	14.21
Cellulase A*		0.95	-0.72	12.09
Cellulase B**		0.99	-0.65	11.46
<b>Cellula</b> se C <sup>†</sup>		0.93	-0.57	11.26
Cellulase $D^{\dagger\dagger}$		0.97	-0.70	11.85
*Purified	d from	Penicillium	C. Notatu	<u>m</u> .

TABLE 4.3.3.--Correlation and Regression Analysis of Activity Data Versus Degree of Polymerization.

\*Purified from Penicillium C. Notatum. \*\*Dialyzed from Aspergillus O. Niger. <sup>†</sup>Partially purified from Aspergillus O. Niger. <sup>†</sup>Purified from Stereum Sanguinolentum.

cellulose. The carboxy-methyl-substituted cellulose was used because samples with a narrow molecular distribution were more readily attainable. The degree of substitution on the cellulose had a range from 0.8 to 1.0. The activity was calculated by relating the changes in viscosity to enzymatic degradation and the enzymatic activity was calculated through the number of bonds broken per unit time; refer to the above paper if additional information on determination of the enzymatic activity is necessary.

The cellulases employed are from three sources: Penicillium Chrysogenim Notatum, Aspergillus Oryzae Niger,


and <u>Stereum Sanguinolentum</u>. These cellulases are all complexes of exo-enzymes and endo-enzymes (116), and random as well as endwise degradation occurs. The relationship between enzymatic activity and degree of polymerization is given in Table 4.3.2 (page 111). A graphical presentation for the <u>Penicillium Notatum</u> complex, illustrating the respective activity versus chain length behavior, is found in Figure 4.3.2 (page 110). The results of correlation and regression analysis (P < 0.0005) on the cellulases' enzymatic activities are given in Table 4.3.3 (page 112).

The activity-degree of polymerization data indicate that the log-linear correlation between DP<sub>j</sub> and a<sub>j</sub> based on equation (4.2.7) is good for both the amyloses and celluloses. This is an important confirmation of my approach, because equation (4.2.7) was derived from an information analysis of carbohydrate encoding, and also is the foundation for subsequent equations relating the information approach to enzyme hydrolysis. The regression analysis yields a considerable variation from the predicted slope of minus one, a necessary condition for maximal information ordering in the system. Only one enzyme, phosphorylase, has a slope close to unity--the others differ. This behavior does not detract from the information-entropy approach but rather implies that these other systems are not most effectively organized. The



carboxymethyl celluloses are certainly not, because substitution of the cellulose polymer is known to have unpredictable effects on the enzyme-substrate reaction (126), which could account for the cellulose-cellulase deviations from unity. The deviation for  $\beta$ -amylase can perhaps be attributed to its mode of enzyme action. Phosphorylase adds glucose monomers to the chain, where  $\beta$ -amylase cleaves off maltose, a glucose dimer; therefore,  $\beta$ -amylase acts on only about half the bonds as would an enzyme cleaving monomers. Thus, a slope between -2 and -1 should be expected because  $\beta$ -amylase's action is relatively quicker. The information-entropy activity relation holds both in the synthesis and degradation of carbohydrate polymers.

The relationship between hydrolysis,  $D_j$ , and degree of polymerization,  $DP_j$ , of carbohydrate molecules <u>in vitro</u>, can be shown for amyloses (127, 128). Four different hydrolyses were conducted on narrowly distributed amylose polymers with  $\beta$ -amylase, using substrates with different initial degrees of polymerization. The results of these experiments are summarized in Table 4.3.4. Hydrolysis,  $D_j$ , is expressed on a scale of 100 instead of 1, and the logarithm of zero is designated as equal to zero. The correlation and regression analysis (see Table 3.3.5) shows a lower degree of correlation and significance than that seen in previous analysis (P < 0.1 for



Test l		Test 2		Test 3		Те	Test 4		Natural Amylose	
DP j	D j	DP j	D.j	DP j	D j	DP j	D j	DPj	D j	
3,150	0%	1,230	0.0%	800	0.0%	795	0.0%	2,600	0.0%	
2,050	20%	730	47.5%	560	29.5%	575	24.5%	2,500	35.5%	
2,110	40%	525	68.0%	350	84.0%	280	78.0%	2,580	53.5%	
1,550	70%	350	91.0%					2,200	72.0%	

TABLE 4.3.4.--Hydrolysis of Amylose Polymers with  $\beta$ -Amylase, and Degree of Polymerization.

Table 4.3.5.--Correlation and Regression Analysis of Hydrolysis and Degree of Polymerization.

	Correlation Coefficient	Slope	y-Intercept
Test l	0.85	-0.14	11.72
Test 2	0.89	-0.22	10.33
Test 3	0.92	-0.17	9.71
Test 4	0.89	-0.21	9.74

test 1 and 2; P < 0.15 for 3 and 4). However, the paucity and limited range of data could considerably bias the results of this analysis. Note that as the range for a particular experiment increases, so does the correlation coefficient. The slopes of all the lines also differ from the theoretical line of minus one.



This deviation of slopes from -1.0 is perhaps best understood after the influence of the reaction order in the encoding process is ascertained. Typically, enzyme reactions are viewed as 1<sup>st</sup> order; substrate and enzyme reacting on a one-to-one basis. However,  $\beta$ -amylase reacts with an average of 4.3 linkages per encounter (129), by a multichain mechanism, yielding a reaction order of 4.3. Such a situation necessitates a revised definition of enzyme activity. If we denote the enzymatic activity for a reaction between the enzyme and one substrate bond as a li, then the total enzymatic activity equals the n<sup>th</sup> product because a<sub>lj</sub> reflects the probability of reaction at one reaction site on the enzyme molecule, and the joint probability that n sites will react determines total enzyme activity and equals the n + 1 product of the activities. Therefore, assuming that the activity at each site is identical, the total enzymatic activity equals:

$$a_{j} = (a_{1j})^{n}$$
 (4.3.1)

The cost function, c<sub>j</sub>, which was initially thought to be equal to the total enzymatic activity, is now seen to be equal to the site activities, a<sub>lj</sub>. Hence, c<sub>j</sub> must be redefined in terms of site activities:

$$c_j = (a_j)^{1/n}$$
, (4.3.2)



which substituted into equation (4.2.7), gives:

$$DP_{j} = \frac{N_{g}}{C_{J}} \frac{1}{(a_{j})^{1/n}} . \qquad (4.3.3)$$

This equation dictates a generalized equation for hydrolysis:

$$k \ln DP_{j} - k \ln DP_{i} = \frac{1}{n} k \ln \left(\frac{a_{i}}{a_{j}}\right)$$
$$= \frac{1}{n} k \ln D_{j} \qquad (4.3.4)$$

Using this result for a reaction order, <u>n</u>, equal to 4.3  $\beta$ -amylase, a slope of -0.232 is expected, closer to the average slope -0.185 which results from experimental data. Why does the previous analysis of DP<sub>j</sub> versus activity not exhibit similar behavior? The reason is that the regression analysis done on the activities listed in Tables 4.3.1 and 4.3.2 was effectively a comparison of the rank orders of the enzymatic activities. The rank ordering of activity for a particular enzyme will be the same for k ln a<sub>1j</sub> or for some constant multiple of it, n k ln a<sub>1j</sub> equal to k ln a<sub>j</sub>. The digestion data were a part of the dynamic analysis of the relative rates of hydrolysis, which are dependent upon reaction order.

The behavior of monodisperse carbohydrate polymers differs from that of natural or polydisperse polymers.

The information analysis can be modified to encompass polydisperse systems. We begin by defining the contribution of the DP<sub>j</sub> polymer fraction to the polydisperse activity of system <u>j</u> as k ln DP<sub>j</sub> times its probable occurrence, P<sub>j</sub>, and do likewise for the DP<sub>i</sub> polymer. Thus, the proportional change in the degree of polymerization of the system is:

$$\sum_{j=1}^{p} k \ln DP_{j} - \sum_{i=1}^{p} k \ln EP_{i}$$

= k ln  $\begin{pmatrix} average \\ chain length \\ ratio \end{pmatrix}$ . (4.3.4)

In Table 4.3.4, the degree of polymerization appears independent of the degree of hydrolysis. Can equation (4.3.4) predict such behavior?

Husemann and Pfannemuller (128) studied the distribution of polymers in polydisperse systems as a function of hydrolysis. Table 4.3.6 presents the chain length distribution of the polydisperse amylose used in the experiment at various stages of hydrolysis. Table 4.3.7 summarizes the above results by giving the expected degree of polymerization, DP<sub>j</sub>, at each stage of hydrolysis for the synthetic and natural amylose. The proportional change is somewhat higher for the synthetic than the natural polymer. The basic chain length behavior of

	0% Digestion		41%		62.5%		85%	
tion	Pj	DP j	Pj	DPj	Pj	DP j	Pj	DP j
1	0.0093	4,950	0.0419	5,000	0.0474	4,800	0.0093	4,900
2	0.0262	4,500	0.0252	4,460	0.0198	4,560	0.1231	4,200
3	0.0322	4,400	0.0315	4,250	0.0937	4,100	0.0526	<b>3,</b> 700
4	0.0193	4,200	0.0557	3,800	0.0651	<b>3,</b> 650	0.1038	3,300
5	0.0438	3,860	0.0842	3,410	0.0367	3,360	0.0980	<b>2,</b> 850
6	0.0229	<b>3,</b> 650	0.0720	<b>3,</b> 020	0.1093	3,000	0.0765	<b>2,</b> 550
7	0.0764	<b>3,</b> 200	0.0774	2,700	0.0887	2,650	0.0666	2 <b>,3</b> 00
8	0.0705	2,850	0.1051	2,350	0.0669	2,350	0.0571	2,100
9	0.0437	2,550	0.0777	2,070	0.0678	2,100	0.0474	1,900
10	0.0712	2,350	0.0684	1,840	0.0419	1,950	0.0565	1,700
11	0.0352	2,200	0.0639	1,650	0.0615	1,750	0.0422	1,030
12	0.0875	1,980	0.0526	1,400	0.0540	1,570	0.0585	1,400
13	0.0748	1,680	0.0526	1,250	0.0421	1,410	0.0451	1,210
14	0.0569	1,420	0.0404	1,100	0.0474	1,230	0.0434	1,080
15	0.0562	1,260	0.0543	940	0.0422	1,100	0.0397	900
16	0.0422	1,120	0.0333	730	0.0251	960	0.0292	725
17	0.0361	970	0.0209	620	0.0323	820	0.0293	580
18	0.0400	840	0.0175	520	0.0137	710	0.0222	430
19	0.0459	700	0.0251	195	0.0120	600		
20	0.0532	420			0.0299	440		
21	0.0572	175						

TABLE 4.3.6.--Chain Length Fractionalization of a Polydisperse Carbohydrate System as a Function of Its Degradation.

Polydisperse Synthetic Amylose					Natu	ural Amy	lose	
& Digestion	Σ <sup>P</sup> j k In DPj	DP. j.*	Chain Length Ratio*	<u></u> **	Chain Length Ratio**	<pre>% Digestion</pre>	<u>DP</u> .	Chain Length Ratio
0	10.660	1,619	1.00	2,180	1.00	0	2,600	1.00
41	10.918	1,934	1.19	2,300	1.06	35.5	2,500	0.96
62.5	11.058	2,131	1.31	2,550	1.17	53.5	2,580	0.99
85	11.011	2,062	1.27	2,350	1.08	72.0	2,220	0.85

TABLE 4.3.7.--Chain Length Behavior of Polydisperse Carbohydrate Systems.

\*Theoretical estimate. \*\*Experimental measurement.

polydisperse systems appears consistent with that predicted by information-entropy analysis.

If the chain length behavior of the system can be predicted by a calculation of the expected chain length, why is the proportional change in degree of polymerization not reflected by the experimental change in hydrolysis? Inspection of the DP<sub>j</sub> data in Table 4.3.6 shows that within each fraction the degree of polymerization remains fairly constant, which can be explained by the constant cascade of higher polymers into lower levels, a result of their degradation. Such analysis does not, however, generate the proper information for application of an information-entropy analysis. What is required is knowledge of the individual behavior of the change in chain length of each amylose polymer in the system, and not the aggregate chain length behavior. Such a study would require uniformly marking the individual polymer (e.g., with a radioactive tracer) so that its degree of polymerization could be studied as its degradation progresses.

In this chapter the viability of informationentropy analyses for rank-ordering enzymatic activities (or encoding costs) of amyloses and celluloses on the bases of their degrees of polymerization has been shown. Also, the utility of this rank-order behavior in calculating the degree of hydrolysis relative to polymer size for monodisperse systems has been put forth. An extension to polydisperse systems was undertaken, and a method for estimating the chain length behavior of these systems was presented. The evidence indicates a great degree of consistency between information-entropy expectations of the encoding behavior of such behavior.



### CHAPTER V

#### DISCUSSION

The goal of this chapter is to provide some additional perspective on the nutritional information-entropy studies undertaken in this work. A review of the main points of each study and of the interrelationships between biological phenomena and model performance will be presented. Also, some discussion of the general ramifications of the information-entropy approach to future studies will be undertaken.

# 5.1 Nitrogen Retention and Information-Entropy

Recall that the model development began with two assumptions: complete digestibility and complete nitrogen retention. An inverse relationship was then seen to exist between two protein quality indices, biological value and net protein value, and rank. These measures were then shown to be related to the entropies of the source and of the decoder. Then the digestibility assumption was dropped and the information-entropy relationship was seen to hold consistently only for net protein value. I now wish to remove the complete nitrogen retention assumption from the model.



As was seen in section 3.2, one of the four metabolic fates of amino acids taken up by the liver is catabolism. The main end product of amino acid catabolism is urea, and the liver accounts for most urea production. The liver, by virtue of its anatomical position, is also the first organ which encounters the amino acid after its encoding into the channel. The most logical route for the catabolic loss of nitrogen appears to be through the liver. It has been experimentally determined for dogs that the liver catabolizes approximately 56% of incoming amino acids, and that the liver is the primary, if not only, organ responsible for the catabolism of essential amino acids (130).

From the information theory viewpoint, this loss of amino acids from the communication system results in a decline of its organizability because each channel is experiencing a loss of informational units. The lack of organizability causes the relative values (i.e., cost) of the system's informational units to change. As you may suspect, this causes an alteration in the expected Zipfian rank-ordering. (Kozachov (58) (see page 29) showed that the measure of organizability in information theory was the regression coefficient, slope, for the rank-frequency function. The organizability is maximum when b/b' equals one, and this condition holds when the information capacity at every level is maximized relative



to the overall information capacity. Therefore, an alteration of amino acid frequency in the system affects the information-entropy level, or alternatively, the information capacity. This is exactly what happens when essential amino acids are catabolized in the liver.

Interpreting the Zipfian slope, we see that when b/b' equals one every j<sup>th</sup> channel information transmission rate equals that of the maximum or rank 1 channel transmission rate, which maximizes the information transfer through the system. As Kozachov would say, the system is maximally organized. A value of b/b' less than one means the j<sup>th</sup> channel transmission rate is not equal to the maximum (rank 1) transmission rate for some j, implying that information is being lost because transmission through the system is less than maximal. For b/b' greater than one the j<sup>th</sup> channel transmission rate is greater than that of rank 1, indicating that the system is exceeding its channel capacity and losing information. Since the system is maximally organized when b/b' is equal to one, the system can be regarded as "underorganized" when b/b' is less than one and "overorganized" when b/b' is greater than one. "Underorganized" systems can be regarded as "leaking" systems since the transmitted information is being lost, while "overorganized" systems can be thought of as "overcompensating" systems because they are



attempting to transmit more information than the system can handle.

A closer inspection allows us to better picture what occurs when the slope deviates from minus one. The mathematical formula for the slope is b/b' where <u>b</u> and <u>b'</u> are proportionality factors between probability and cost (see page 28), such that:

$$P_{j} = \exp(-bc_{j})$$
 (5.1.1)

and

$$P_{1} = \exp(-b'c_{1}) . \qquad (5.1.2)$$

Solving equations (5.1.1) and (5.1.2) for b and b', we get:

$$b/b' = \frac{\ln P_j}{c_j} \cdot \frac{c_l}{\ln P_l}$$
 (5.1.3)

Now, the j<sup>th</sup> channel transmission rate is:

$$I_{j} = \frac{\ln (n_{j})}{t_{j}} = \frac{-\ln P_{j}}{c_{j}}.$$
 (5.1.4)

Therefore, the slope b/b' may be expressed in terms of channel transmission rates:

$$b/b' = \frac{\ln P_j}{c_j} \cdot \frac{c_1}{\ln P_1} = \frac{I_j}{I_1}$$
 (5.1.5)

the second s



A deviation from maximum organizability was noted in the Zipfian analyses conducted in section 3.3, where the slopes differed greatly from unity. Both the "underorganized" and "overorganized" systems are present. Since the organizability of the protein information system is dependent upon the amounts of amino acids flowing through the channels, an underorganized system having b/b' values less than one reflects the loss of organization in the system resulting from catabolism. The overorganized systems, those with b/b' greater than one, reflect the conservation of amino acids in the protein information system; that is, the information transmission rate of the j<sup>th</sup> channel is increased so more information can pass through this channel.

Introduction of the catabolic concept into our system effectively allows rejection of the complete retention hypothesis. The Zipf's law equation for a single amino acid becomes:

 $k \ln ax_{j} = k \ln am_{j} - f_{c} k \ln r_{xj}$ , (5.1.6)

or, for the entire essential amino acid set:

$$\overline{H}_{x}(EAA) = \overline{H}_{m}(EAA) - f_{c} k \ln r_{x}$$
, (5.1.7)

where  $r_x$  is the rank of protein  $\underline{x}$ ,  $\overline{H}_m$  (EAA) is the rank one log-frequency, and  $f_c$  is a catabolism factor



accounting for the loss of organization in the system due to amino acid destruction by the liver.

The regression analyses done in section 3.3 give, for the information-entropy measures based on the essential amino acid set, a range for  $\underline{f_C}$  of -0.517 to -0.608 for rats, and -0.601 to -0.700 for pigs. The interpretation of this result is that between 40% and 50% of the essential amino acids in the diets of rats are catabolized, and between 30% and 40% of those in the diets of baby pigs are converted to urea. Both of these values are comparable to the experimentally determined value for mature dogs of approximately 56%. Younger animals such as the experimental pigs and rats could be expected to have a higher nitrogen retention.

If these urea production figures for baby pigs and rats are correct, a new use for the informationentropy model has arisen. In addition to being able to predict protein quality indices, the model can in turn be used to estimate the degree of catabolism of essential amino acids for various species of animals. The experimental method for measuring this phenomenon is very difficult, and the information-entropy model may quite possibly provide an adequate alternative.

The interpretation of an overorganized system as overcompensating and thus reflecting essential amino acid conservation is supported by the regression analyses



in section 3.3. A formula similar to equation (5.1.7) can be employed to reflect amino acid conservation:

$$\min H_x(aa_j) = \min H_s(aa_j) - f_o \cdot k \ln r_{xj}, \quad (5.1.8)$$

where  $f_0$  is a conservation factor which represents the overcompensation in the system due to essential amino acid conservation, and min  $H_s(aa_j)$  is the log frequency for the standard protein.

Returning to the regression analyses in section 3.3, the range for  $f_0$  for rats is -1.287 to -1.487, while the range for pigs is -1.097 to -1.460. These results imply a conservation of the limiting essential amino acid. The degree to which it is conserved is not readily evident: whereas the slope for underorganized systems can range from 0.0 to 1.0, that for overorganized systems goes from 1.0 to infinity. Nonetheless, the informationentropy model conforms to the rule that the most limiting amino acid is conserved by the organism.

The question now arises as to how the inclusion of  $\underline{f_{C}}$  will affect the ability of the information-entropy model to predict net protein value, the protein quality index. The mathematics involved are quite simple and for the essential amino acid set, the logarithm of NPV, NPV<sub>x</sub>(EAA) is:

NPV<sub>x</sub>(EAA) = 
$$\frac{1}{f_c} \bar{H}_x(EAA) - \frac{1}{f_c} \bar{H}_s(EAA)$$
. (5.1.9)



The predicted NPV resulting from utilizing  $f_c$  will be denoted  $I_x$  (EAAR) and known as the "essential amino acid retention index":

$$I_{x}(EAAR) = antilog \frac{1}{f_{c}} \vec{H}_{x}(EAA) - \frac{1}{f_{c}} \vec{H}_{s}(EAA)$$
(5.1.10)

when based on log-frequency  $\bar{H}_{x}(EAA)$ , or  $I_{x}^{0}(EAAR)$  for log-frequency  $\bar{H}_{x}^{0}(EAA)$ .

Utilizing equation (5.1.10) with  $f_c = 0.5$ ,  $I_x(EAAR)$  and  $I_x^0(EAAR)$  were determined from the data in section 3.3. Table 5.1.1 lists the correlation coefficients among  $I_x(EAAR)$  and  $I_x^0(EAAR)$ , and the experimental net protein values for rats and pigs. Table 5.1.2 lists the slopes of a linear regression analysis for the above variables and Table 5.1.3 lists the corresponding y-intercept values.

The correlation coefficients are not very different from those obtained by regression without  $f_c$ . However, the regression analysis shows a much improved picture in the ability of the model to accurately predict the true score of net protein value for rats but the results for pigs indicate an  $f_c = 0.5$  may be too high a catabolism factor. This is graphically illustrated in Figures 5.1.1 and 5.1.2 for rats and pigs, respectively. These graphs show the relationships of the model's information-entropy measures,  $I_x^0$ (EAA) and  $I_x^0$ (EAAR), to

		Net Protein Value (Rats)	Net Protein Value (Pigs)
FAO Data:	I <sub>x</sub> (EAAR <sub>f</sub> )	0.854	0.806
	$I_x^0(EAAR_f)$	0.879	0.826
Eggum Data:	I <sub>x</sub> (EAAR <sub>e</sub> )	0.851	0.819
	$I_x^0$ (EAAR <sub>e</sub> )	0.910	0.842

TABLE 5.1.1.--Correlation Coefficients Among Essential Amino Acid Retention Indices and Experimental Protein Values of Rats and Pigs.

TABLE 5.1.2.--Linear Regression Coefficients Among the Essential Amino Acid Retention Indices and Experimental Net Protein Values of Rats and Pigs.

		Net Protein Value (Rats)	Net Protein Value (Pigs)
FAO Data:	I <sub>x</sub> (EAAR <sub>f</sub> )	0.642	0.533
	$I_x^0$ (EAAR <sub>f</sub> )	0.657	0.543
Eggum Data:	I <sub>x</sub> (EAAR <sub>e</sub> )	0.678	0.573
	$I_x^0$ (EAAR <sub>e</sub> )	0.800	0.650

TABLE 5.1.3.--Slopes for Regression Analysis Among Essential Amino Acid Indices and Experimental Net Protein Values of Rats and Pigs.

	han an a	Net Protein	Net Protein
FAO Data:	I <sub>x</sub> (EAAR <sub>f</sub> )	23.7	35.5
	$I_x^0(EAAR_f)$	24.0	35.8
Eggum Data:	I <sub>x</sub> (EAAR <sub>e</sub> )	20.7	32.3
	$I_x^0(EAAR_e)$	15.1	29.1



Figure 5.1.1.--Graph of Information-Entropy Indices  $I_x^0$  (EEAR<sub>e</sub>) and  $I_x^0$ (EAA<sub>e</sub>) Versus Net Protein Value for Rats (Source: Eggum).







the experimental net protein values. It is readily seen that for rats there is a much better one-to-one correspondence for the Essential Amino Acid Retention index than for the traditional Essential Amino Acid Index, for the latter does not account for liver catabolism of amino acids. However, the shift of the data points on the pig graph suggest that a lower  $f_c$  would yield a better oneto-one correspondence.

The information-entropy model is thus capable of predicting indices of protein quality and other physiological behavior associated with protein metabolism (e.g., urea production and conservation of the limiting amino acid). The utilization of the information-entropy model is justified, I believe, by such performance, and holds promise for still other applications.

## 5.2 The Information-Entropy Model for Protein Metabolism: Summary

In this section, I wish briefly to present the major mathematical relationships utilized in my information-entropy model. First, it should be recognized that the model works on two levels. The first level is that of a single amino acid channel. For the single channel model, the outcome-generating form is:

$$k \ln NPV(x_j) - k \ln NPV(s_j) = H_x(aa_j) - H_s(aa_j)$$
. (5.2.1)


The significance of the single channel model is that when protein <u>s</u> possesses a net protein value of one, the essential amino acid channel which minimizes equation (5.2.1) is taken to predict the net protein value of protein <u>x</u>. This is a method identical to the chemical scoring of amino acids.

The other level used by the model is the multichannel level. Here, an average of the informationentropy levels of the essential amino acid channels generates the outcome. The following equation is employed for multichannel calculations:

$$NPV_{x}(EAA) - NPV_{s}(EAA) = \bar{H}_{x}(EAA) - \bar{H}_{s}(EAA) . \qquad (5.2.2)$$

The multichannel form is also used to predict the net protein value of food protein  $\underline{x}$  when the NPV of protein  $\underline{s}$  is known. If the net protein value of protein  $\underline{s}$  is one, and Oser's rules are employed, the method is the same as calculating the essential amino acid index.

The multichannel model was adapted in the previous section to account for catabolism of essential amino acids by the liver. This involved introducing  $\underline{f_c}$ , the uncatabolized fraction of amino acids. The mathematical form for this application of the informationentropy approach is:

$$NPV_{x}(EAAR) - NPV_{s}(EEAR) = \frac{1}{f_{c}} \overline{H}_{x}(EAA) - \frac{1}{f_{c}} \overline{H}_{s}(EAA).$$
 (5.2.3)

,

Once again, we assume net protein value to be the index of protein quality predicted by the antilogarithmic form of equation (5.2.3) when the NPV of protein <u>s</u> is known. This formulation is not similar to any measure of protein quality based upon the amino acid composition of the food proteins involved. It has been termed the "essential amino acid retention index" and may be determined with or without Oser's rules. A value of  $f_c$  equal to 0.5 may be assumed for the rat, the standard test animal.

This discussion presents the major functional forms of the model. With the appropriate assumptions, we can obtain any index of protein quality given in this thesis.

# 5.3 Information-Entropy and Polymers: An Appraisal

The use of an entropic concept to study polymer behavior is most logical. Entropy is usually associated with order, and the chain length of a polymer is one of the most obvious examples of order on a molecular scale. This study, like the previous one, emphasizes the importance of cost-frequency behavior. The most important development was the derivation of the idea that the cost of hydrolysis of the polymer message was equal to the inverse of the enzyme activity.

This result from information-entropy analysis can be confirmed by the Michaelis-Menten enzyme kinetic model. The reaction velocity, v, is defined:

$$v = a_j (enzyme activity) = \frac{V_m S'}{K_m + S} (moles/sec.)$$
 (5.3.1)

where S' is substrate concentration, which for a monodisperse polymer solution, may be defined  $N_g/DP_j$ . If  $K_m << S'$ , then  $a_j$  equals  $V_m$  and enzyme activity is invariant with change in degree of polymerization. From the information-entropy perspective, this condition overloads the encoding channel, exceeding its capacity, and because of this the encoding device does not exhibit cost-frequency behavior.

Alternatively, if  $K_m >> S'$ , then equation (5.3.1) has the following form:

$$a = \frac{V_{m}}{K_{m}} S' = \frac{V_{m}}{K_{m}} \frac{N_{g}}{DP_{j}}, \qquad (5.3.2)$$

or

This is the exact conclusion of the information-entropy analysis. Thus, a valid analog to the encoding model presented in this chapter can be found in a special application of enzyme kinetics. Extension of the model to predict digestibilities had questionable results. From



my viewpoint, this was due to the lack of adequate data. However, the regression analysis did not indicate that the log-log relationship between degree of polymerization and digestibility had complete merit. Rather, the level of significance of the results did not offer sufficient cause for acceptance.

The most pertinent notion gleaned from this analysis is the inverse proportionality between enzyme activity and degree of polymerization. It indicates that introduction of an information-entropy formalism may be possible for the study of enzyme kinetics.

## CHAPTER VI

### CONCLUSIONS

1. The information-entropy model of protein metabolism can be employed to assess the nutritional quality of proteins. This is accomplished by relating the amino acid content of proteins in the diet to net protein value. The model's output is similar to other amino acid scoring approaches such as Oser's essential amino acid index and chemical score.

2. A new index, the "essential amino acid retention index," was postulated from the information-entropy model. It was as well correlated to net protein value as other chemical scoring methods, and permitted the catabolism of ingested amino acids to be accounted for.

3. An information-entropy analysis of polymer length versus the activity of enzymatic hydrolysis, by a cost-frequency analysis of encoding, was well correlated with experimental data on the subject.

4. The extension of the information-entropy study for the estimation of a hydrolysis coefficient could not be adequately correlated with experimental data.



REFERENCES

.



#### REFERENCES

- (1) S. Carnot, <u>Reflections on the Motive-power of Heat</u>, 1824, trans. by R. H. Thurston, ed. by E. Mendoza (New York: Dover Publications, 1960), p. 7.
- (2) R. Clausius, <u>The Mechanical Theory of Heat</u> (London: Macmillan and Co., 1879), p. 78.
- (3) W. Thomson (Kelvin), "On the Dynamical Theory of Heat with Numerical Results Deduced from Mr. Joule's Equivalent of a Thermal Unit, M. Regnault's Observations on Steam," Trans. of the Royal Society of Edinburgh (March, 1851), reprinted in The Second Law of Thermodynamics, ed. by W. F. Magie (New York: Harper & Brothers, 1899), pp. 111-148.
- (4) J. C. Maxwell, Theory of Heat (10th ed.; London: Longmans, Green & Co., 1921), p. 189.
- (5) C. Caratheodory, Math. Ann., No. 67 (1909), p. 355, cited by S. Blinder, "Caratheodory's Formulation of the Second Law," Physical Chemistry, ed. by W. Dost (New York: Academic Press, 1971), pp. 613-637.
- (6) M. J. Klein, "The Development of Boltzman's Statistical Ideas," <u>The Boltzman Equation</u>, ed. by E. G. D. Cohen and W. Thirring (New York: Springer-Verlag, 1973).
- L. Boltzman, "Further Studies on the Thermal Equilibrium of Gas Molecules," <u>Wiener Berichte</u>, v. 66 (1872), cited by L. Boltzman, <u>Lectures on Gas</u> <u>Theory</u> (1896), trans. by S. Brush (Berkeley: University of California Press, 1964), p. 52.
- (8) J. C. Maxwell, <u>Matter and Motion</u> (New York: Dover Publications, Inc., 1877).
- (9) L. Boltzman, "Observations on One Problem of the Mechanics of Heat Theory," <u>Wiener Berichte</u>, v. 76 (1877), cited in L. Boltzman, <u>Lectures on Gas</u> <u>Theory</u> (1910), trans. by S. Brush (Berkeley: University of California Press, 1964), p. 58.



- (10) G. Arfken, <u>Mathematical Methods for Physicists</u> (New York: Academic Press, 1965).
- (11) J. W. Gibbs, <u>Elementary Principles in Statistical</u> <u>Mechanics</u> (New Haven: Yale University Press, 1902).
- (12) E. H. Kerner, <u>Gibbs Ensemble: Biological Ensemble</u> (New York: Gordon and Breach, Science Publishers, 1972), p. vii.
- (13) J. Kestin and J. R. Dorfman, <u>A Course in Statistical</u> <u>Thermodynamics</u> (New York: Academic Press, 1971), pp. 178-181.
- (14) Ibid., p. 196.
- (15) Ibid., p. 199.
- (16) E. T. Jaynes, "Gibbs vs. Boltzman Entropies," Amer. J. Physics, v. 33 (1965), pp. 391-398.
- (17) A. Grunbaum, "Is the Coarse-grained Entropy of Classical Statistical Mechanics an Anthropomorphism," Modern Developments of Thermodynamics, ed. by B. Gallor (New York: J. Wiley & Sons, 1974), pp. 413--28.
- (18) L. Szilard, "On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings," <u>Z. Phy.</u>, v. 53 (1929), trans. in <u>Behavioral</u> <u>Science</u>, v. 9 (October, 1964), pp. 301-310.
- (19) C. E. Shannon, "A Mathematical Theory of Communication," <u>Bell System Tech. J.</u>, v. 27 (July-October, 1948), pp. 379 and 623.
- (20) J. R. Pierce, "The Early Days of Information Theory," <u>IEEE Trans. on Info. Thy</u>., v. IT-19, No. 1 (January, 1973).
- (21) R. V. L. Hartley, "Transmission of Information," Bell System Tech. J., v. 7 (July, 1928), pp. 535-563.
- (22) D. Gabor, "New Possibilities in Speech Transmission," J. Inst. Elect. Eng. (London), v. 94 (November, 1947), pp. 369-390.
- (23) A. I. Khinchin, <u>Mathematical Foundations of Informa-</u> tion Theory (New York: Dover Publications, 1957), pp. 9-13.



- (24) E. T. Jaynes, "Information Theory and Statistical Mechanics," <u>Physical Review</u>, v. 106, No. 4 (1957), pp. 620-630.
- (25) E. T. Jaynes, "Information Theory and Statistical Mechanics, II," <u>Physical Review</u>, v. 108, No. 2 (1957).
- (26) M. Tribus, P. T. Shannon, and R. B. Evan, "Why Thermodynamics is a Logical Consequence of Information Theory," AIChE J. (March, 1966), p. 244.
- (27) P. Ziesche, "About a New Introduction of Entropy in Statistical Mechanics due to Macke," <u>Proc. of</u> <u>Coll. on Info. Thy.</u>, v. II, ed. by A. Renyi (Budapest, Hungary: J. Bolyai Math. Soc., 1968), pp. 515-518.
- (28) J. Fritz, "Information Theory and Thermodynamics of Gas Systems," Proc. of Coll. on Info. Thy., v. I, ed. by A. Renyi (Budapest, Hungary: J. Bolyai Math. Soc., 1968), pp. 167-175.
- (29) M. Tribus, <u>Thermostatics and Thermodynamics: An</u> <u>Information Theory Approach</u> (Princeton, N.J.: Van Nostrand, 1961).
- (30) R. Baierlain, Atoms and Information Theory (San Francisco: Freeman and Company, 1971).
- (31) D. C. Zubaren, <u>Nonequilibrium Statistical Thermodynamics</u> (New York: Consultants Bureau, 1974), pp. 38-45 and 100-103.
- (32) J. von Neumann, <u>Theory of Self-Reproducing Automata</u> (Urbana, Illinois: University of Illinois Press, 1966), pp. 60-61.
- (33) G. Jumarie, "Further Advances on the General Thermodynamics of Open Systems via Information Theory: Effective Entropy, Negative Information," Int. J. of Sys. Sci., v. 6, 1975, pp. 249-269.
- (34) I. N. Taganov, "Information Simulation of Multifactor Systems in Chemistry and Chemical Engineering," <u>Theoretical Foundations of Chemical Engineer-</u> ing, English translation from the Russian, v. 9, No. 2 (1975, translated January, 1976), pp. 223-228.



- (35) D. Slepian, "Information Theory in the Fifties," <u>IEEE Trans. on Info. Thy.</u>, v. IT-19, No. 2 (March, 1973), pp. 143-148.
- (36) C. Cherry, ed., Information Theory: 3rd London Symposium (London: Burtersworths, 1956).
- (37) R. W. Hamming, "Error Detecting and Error Correcting Codes," <u>Bell System Tech. J.</u>, v. 29 (1950), pp. 147-160.
- (38) C. E. Shannon, "Certain Results in Coding Theory for Noisy Channels," <u>Inform. Control</u>, v. 1 (September, 1957), pp. 6-25.
- (39) A. J. Viterbi, "Information Theory in the Sixties," <u>IEEE Trans. on Info.</u>, v. IT-19, No. -3 (May, 1973), pp. 257-262.
- (40) E. R. Berlekamp, Algebraic Coding Theory (New York: McGraw-Hill, 1968).
- (41) J. K. Wolf, "A Survey of Coding Theory: 1967-1972," <u>IEEE Trans. on Info. Thy</u>., v. IT-19, No. -4 (July, 1973), pp. 381-389.
- (42) W. R. Garner, <u>Uncertainty and Structure as Psycho-logical Concepts</u> (New York: John Wiley and Sons, Inc., 1962), pp. 28-32.
- (43) H. Theil, <u>Economics and Information Theory</u> (Amsterdam, Netherlands: North-Holland Publishing Co., 1967).
- (44) J. M. Cozzolino and M. J. Zahner, "The Maximum-Entropy Distribution of Future Market Price of Stock," <u>Operation Research</u>, v. 21 (1973), pp. 1200-1211.
- (45) B. Lev, "Accounting and Information Theory," <u>Studies</u> <u>in Accounting Research</u> (American Accounting Association, 1969).
- (46) N. Georgescu-Roegen, The Entropy Law and the Economic Process (Cambridge, Massachusetts: Harvard University Press, 1971).
- (47) L. W. Rosenfield, Aristotle and Information Theory (The Hague, Netherlands: Humanities Press, 1971).

- (48) M. A. P. Willmer, "Information Theory and the Measurement of Detective Performance," <u>Kybernetes</u>, v. 2 (1973), pp. 225-231.
- (49) L. L. Gatlin, "The Entropy Maximum of Protein," Math. Biosci., v. 13 (1972), pp. 213-227.
- (50) M. Haegawa and T. Yanko, "The Genetic Code and the Entropy of the Protein," <u>Math. Biosci.</u>, v. 24 (1975), pp. 169-182.
- (51) C. E. Shannon, "Prediction and Entropy of Printed English," <u>Bell System Tech. J.</u>, v. 30 (1951), p. 50.
- (52) J. F. Young, <u>Information Theory</u> (New York: Wiley-Interscience, 1971), pp. 50-58.
- (53) C. E. Shannon and W. Weaver, <u>The Mathematical</u> <u>Theory of Communication</u> (1st paperback ed.; Urbana, <u>Illinois:</u> University of Illinois Press, 1949).
- (54) L. P. Hyvarinen, <u>Information Theory for Engineers</u> (New York: Springer-Verlag, 1968), pp. 15-17.
- (55) B. Mandelbrot, Jeux de communication (Institut de Statistique de l'Université de Paris, 1953), cited by L. Brillouin, Science and Information Theory (New York: Academic Press, Inc., 1962), pp. 28-47.
- (56) G. K. Zipf, The Psycho-Biology of Language (Cambridge, Massachusetts: MIT Press, 1935).
- (57) G. K. Zipf, <u>Human Behavior and the Principle of</u> <u>Least Effort</u> (Reading, Massachusetts: Addison-Wesley Press, Inc., 1949).
- (58) L. S. Kozachkov, "Certain Integral Properties of Information Systems of Hierarchic Type," <u>Kiber</u>netica (1974).
- (59) V. T. Coates, <u>Revitalization of Small Communities</u>: <u>Transportation Options</u>, U.S. Department of Transportation, DOT-TST-75-1 (May, 1974).
- (60) V. Pareto, <u>Manual of Political Economy</u>, trans. by A. S. Schwier, ed. by A. S. Schwier and A. N. Page (New York: A. M. Kelley, 1971).



- (61) J. Lotka, "The Frequency Distribution of Scientific Productivity," J. Acad. Sci. (Washington, D.C., No. 12, 1926).
- (62) J. Huxley, Problems of Relative Growth (2nd ed.; New York: Dover Press, 1972).
- (63) T. A. Loomis, Essentials of Toxicology (Philadelphia: Lea & Febiger, 1968).
- (64) M. Tribus, <u>Rational Descriptions</u>, <u>Decisions and</u> Designs (New York: Pergamon Press, 1969).
- (65) E. T. Jaynes, Probability Theory in Science and Engineering (Dallas, Texas: Mobil Oil Research Laboratory, 1959).
- (66) R. T. Cox, <u>The Algebra of Probable Inference</u> (Baltimore, <u>Maryland</u>: Johns Hopkins Press, 1961), pp. 35-65.
- (67) C. E. Shannon, "The Bandwagon," <u>IEEE Trans. Info.</u> Thy., v. IT-2 (March, 1956), p. 3.
- (68) H. P. Yockey, R. L. Platzman and H. Quastler, editors, Symposium on Information Theory in Biology (New York: Pergamon Press, 1958).
- (69) S. M. Danoff and H. Quastler, editors, Essays on the Use of Information Theory in Biology (Urbana, Illinois: University of Illinois Press, 1953).
- (70) W. M. Elsasser, <u>The Physical Foundations of Informa-</u> tion Theory in Biology (New York: Pergamon Press, 1958).
- (71) E. Samuel, Order: In Life (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972).
- (72) J. von Neumann, "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components," <u>Ann. of Math. Studies</u>, No. 34 (1956), pp. 43-98.
- (73) B. Michaels and R. A. Chaplain, "The Encoder Mechanism of Receptor Neurons," <u>Kybernetic</u>, v. 13 (1973), pp. 6-23.
- (74) M. Abeles, "Transmission of Information by the Axon: II--The Channel Capacity," <u>Biological Cybernetics</u>, v. 19 (1975), pp. 121-125.

----

- (75) R. F. Quick and T. A. Reichert, "Multi-Channel Models of Human Vision: Bandwidth Considerations," Kybernetic, v. 12 (1973), pp. 141-144.
- (76) A. B. Kogan and O. G. Chorajan, "Some Information Theory Applications to the Physiology of the Nervous Cell," Kybernetes, v. 2 (1973), pp. 77-78.
- (77) R. Eckhorn and B. Popel, "Rigorous and Extended Application of Information Theory to the Affernet Visual System of the Cat-II: Experimental Results," Biological Cybernetics, v. 17 (1975), pp. 7-17.
- (78) M. W. Nirenberg and H. Matthaei, "Dependence of Cell-Free Protein Synthesis in <u>E. coli</u> upon Naturally Occurring or Synthetic Polyribonucleotides," <u>Proc. Nat. Acad. Sci.</u> (USA), v. 47 (1961), p. 1588.
- (79) H. R. Mahler and E. H. Cordes, <u>Biological Chemistry</u> (New York: Harper and Row, 1971), p. 783.
- (80) L. L. Gatlin, Information Theory and the Living System (New York: Columbia University Press, 1972).
- (81) L. M. Spetner, "Information Transmission in Evolution," <u>IEEE Transactions on Information Theory</u>, v. IT-14, No. 1 (1968), pp. 3-6.
- (82) E. Schröedinger, What is Life? (Cambridge, England: Cambridge University Press, 1945).
- (83) J. von Neumann, Theory of Self-Reproducing Automata (Urbana, Illinois: University of Illinois Press, 1966), p. 61.
- (84) P. Fong, "Thermodynamic and Statistical Theory of Life: An Outline," <u>Biogenesis</u>, Evolution, Homeostasis--A Symposium by Correspondence, ed. by A. Locker (Berlin, Germany: Springer-Verlag, 1975), pp. 93-100.
- (85) P. Glansdorff and I. Prigogine, <u>Thermodynamic Theory</u> of Structure, Stability and Fluctuations (New York: Wiley-Interscience, 1971).
- (86) L. A. Maynard and J. K. Loosli, <u>Animal Nutrition</u> (New York: McGraw-Hill Book Co., 1969), p. 367.
- (87) M. Hawegawa and Y. Taka-aki, "The Genetic Code and the Entropy of a Protein," <u>Mathematical Biosciences</u>, v. 24 (1975), pp. 169-182.



- (88) K. Thomas, "Uber die Biolische Wertigkeit der Stickstoff Substanzen in ver schieden Nahrungsmittel," Arch. Anat. u. Physiol., Physiol. Abstract (1909), pp. 212-302, cited by Maynard and Loosli, op. cit., p. 459.
- (90) H. H. Mitchell and G. G. Carman, "The Biological Value of the Protein Nitrogen Mixtures of Patent White Flour and Animal Foods," J. Biol. Chem., v. 68 (1926), pp. 183-215.
- (91) A. E. Bender and D. S. Miller, "A Brief Method for Estimating the Value of Protein," <u>Biochem. J.</u>, v. 53 (1953), p. vii.
- (92) D. S. Miller and A. E. Bender, "The Determination of the Net Utilization of Proteins by a Shortened Method," <u>British J. of Nutrition</u>, v. 9 (1955), pp. 382-390.
- (93) J. M. McLaughlan and J. A. Campbell, "Methodology of Protein Evaluation," <u>Mammalian Protein Metabolism</u>, <u>Vol. III</u>, ed. by H. N. Munro (New York: Academic Press, 1969), pp. 391-422.
- (94) T. B. Osborne, L. B. Mendel and E. L. Perry, "A Method of Expressing Numerically the Growth-Promoting Value of Proteins," J. Biological Chem., v. 37 (1919), p. 223.
- (95) D. V. Frost, "Methods of Measuring the Nutritive Value of Proteins, Protein Hydrolyzates, and Amino Acid Mixtures," <u>Protein and Amino Acid Nutrition</u>, ed. by A. A. Albanese (New York: Academic Press, 1959), pp. 225-274.
- (96) D. M. Hegsted, "Assessment of Protein Quality," <u>Improvement of Protein Nutriture</u>, National Academy of Sciences (1974), pp. 64-88.
- (97) Amino Acid Content of Foods and Biological Data on <u>Proteins</u>, FAO: FAO Nutritive Studies, Rome, Italy (1970), No. 24.
- (98) S. B. Richmond, <u>Statistical Analysis</u> (2nd ed.; New York: Ronald Press Co., 1964), pp. 424-465.
- (99) R. J. Senter, <u>Analysis of Data</u> (Glenview, Illinois: Scott, Foresman, & Co., 1969), pp. 440-445.



- (100) H. H. Mitchell and R. J. Block, "Some Relationships Between the Amino Acid Contents of Proteins and their Nutritional Values for the Rat," <u>J. Biol.</u> <u>Chem.</u>, v. 163 (1946), p. 599.
- (101) B. L. Oser, "Method for Integrating Essential Amino Acid Content in the Nutritional Evaluation of Proteins," J. Amer. Dietetic Assoc., v. 27 (1951), pp. 396-402.
- (102) A. E. Bender, "Rat Assays for Protein Quality--A Reappraisal," Proc. 9th International Congress on Nutrition (Mexico City, Mexico: 1972), v. 3, reprinted by Karger Basel (1975), pp. 310-320.
- (103) H. N. Munro, "A General Survey of Mechanisms Regulating Protein Metabolism in Mammals," <u>Mammalian</u> <u>Protein Metabolism</u>, ed. by H. N. Munro (New York: <u>Academic Press, 1970</u>), v. 4, pp. 3-130.
- (104) A. M. Rosie, Information and Communication Theory
  (London, England: Van Nostrand Reinhold Co., 1973),
  p. 90.
- (105) H. H. Williams, A. E. Harper, D. M. Hegsted, <u>et al.</u>, "Nitrogen and Amino Acid Requirements," <u>Improvement</u> <u>of Protein Nutriture</u>, National Academy of Sciences (1974), pp. 23-63.
- (106) H. N. Munro, "Free Amino Acids and Their Role in Regulation," <u>Mammalian Protein Metabolism</u>, v. 4, ed. by H. N. Munro (New York: Academic Press, 1970), pp. 299-386.
- (107) J. M. McLaughlan and A. B. Morrison, "Dietary Factors Affecting Plasma Amino Acid Concentrations," <u>Protein Nutrition and Free Amino Acid Patterns</u>, ed. by J. H. Leathem (New Brunswick, New Jersey: Rutgers University Press, 1968), pp. 3-18.
- (108) C. Gitler, "Protein Digestion and Absorption in Non-Ruminants," <u>Mammalian Protein Metabolism</u>, v. 1, ed. by H. N. Munro (New York: Academic Press, 1964), pp. 35-70.
- (109) D. H. Elwyn, "Modification of Plasma Amino Acid Pattern by the Liver," <u>Protein Nutrition and Free</u> <u>Amino Acid Patterns</u>, ed. by J. H. Leathem (New Brunswick, New Jersey: Rutgers University Press, 1968), pp. 88-106.



- (110) A. N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information," <u>Problems of</u> <u>Information Transmission</u>, v. 1, No. 1 (1965), pp. 3-11.
- (111) B. O. Eggum, <u>A Study of Certain Factors Influencing</u> <u>Protein Utilization in Rats and Pigs</u> (Copenhagen, <u>Denmark: I Kommission has Lanhusholdningsselskabets</u> Forlag, 1973).
- (112) I. Danishefsky, R. L. Whistler and F. A. Bettelheim, "Introduction to Polysaccharides," <u>The Carbohy-</u> <u>drates, Chemistry and Biochemistry</u>, ed. by W. Pigman and D. Horton (New York: Academic Press, 1970), v. II-A, pp. 375-410.
- (113) C. T. Greenwood, "Starch in Glycogen," ed. by W. Pigman and D. Horton, "<u>The Carbohydrates, Chemistry</u> and Biochemistry (New York: Academic Press, 1970), v. II-B, pp. 471-513.
- (114) E. B. Cowling, "Structural Features of Cellulose," ed. by E. T. Reese, Advances in the Enzymatic Hydrolysis of Cellulose and Related Materials (New York: Pergamon Book Press, 1963).
- (115) W. S. Whelan, "Enzymatic Explorations of the Structures of Starch and Glycogen," <u>Biochemistry</u>, v. 122 (1971), pp. 609-622.
- (116) K. W. King, "Enzymes of the Cellulase Complex," ed. by R. F. Gould, <u>Cellulases and Their Applica-</u> <u>tions</u>, Advances in Chemistry Series 95 (Washington, <u>D.C.</u>: American Chemical Society Publications, 1969), pp. 7-26.
- (117) A. White, P. Handler, and E. Smith, <u>Principles of</u> <u>Biochemistry</u> (New York: McGraw-Hill, Inc., 1973), pp. 412-504.
- (118) Weende Method, cited by L. A. Maynard and J. K. Loosli, Animal Nutrition (New York: McGraw-Hill, Inc., 1969), pp. 76-77.
- (119) <u>Biological Energy Interrelationships and Glossary</u> <u>of Energy Terms</u>, National Academy of Sciences (Washington, D.C.: Printing and Publishing Office, NAS, 1966), Publication No. 1411.
- (120) R. A. Gibbons, Polydispersity," <u>Nature</u>, v. 200 (November 16, 1963), pp. 665-666.

- (121) F. M. Reza, <u>An Introduction to Information Theory</u> (New York: <u>McGraw-Hill</u>, 1961), pp. 132-135.
- (122) H. R. Mahler, E. H. Cordes, <u>Biological Chemistry</u> (New York: Harper & Row, 1966), p. 230.
- (123) L. Michaelis and M. L. Menten, <u>Biochem. Z.</u>, v. 49 (1913), p. 333.
- (124) E. Husemann and B. Pfannemuller, "An Investigation of the Kinetics of  $\beta$ -Amylase and Phosphorylase: The Dependence of Reaction Velocity on the Chain Length of the Amylose," D. Makromole. Chem., v. 87 (1965), pp. 139-151.
- (125) W. Z. Hassid, "Biosynthesis of Sugars and Polysaccharides," <u>The Carbohydrate's Chemistry and</u> <u>Biochemistry</u>, ed. by W. Pigman and D. Horton (New York: Academic Press), v. II-A, pp. 302-373.
- (126) K. E. Almin and K. E. Eriksson, "Influence of Carboxymethyl Cellulose Properties on the Determination of Cellulase Activity in Absolute Terms," Archiv. Biochem. Biophys., v. 124 (1968), pp. 129-134.
- (127) E. Husemann and B. Pfannemuller, "On the Distinctive Behavior of Synthetic and Natural Amylose in Comparison with Phosphorylase and β-Amylase,"
   D. Makromole. Chem., v. 83 (1961), p. 157.
- (128) E. Husemann and B. Pfannemuller, "On the Degradation of Synthetic and Natural Amylose by Potato Phosphorylase and  $\beta$ -Amylase II," <u>D. Makromole. Chem.</u>, v. 85 (1963), pp. 74-95.
- (129) J. M. Bailey and D. French, "The Significance of Multiple Reactions in Enzyme-Polymer Systems," J. Biol. Chem., v. 226 (1957), p. 1.
- (130) D. H. Elwyn, "The Role of the Liver in Regulation of Amino Acid and Protein Metabolism," <u>Mammalian</u> <u>Protein Metabolism</u>, v. 4, ed. by H. N. Munro (New York: Academic Press, 1970), pp. 523-558.

# 

÷





