THE LINGUISTIC GENETIC MESSAGE

ANTHONY SEBASTIAN
San Francisco, California

The Proper Study of Mankind

In his Essay on Man, Alexander Pope exhorted, "Know then thyself, presume not God to scan; / The proper study of Mankind is Man. / ...The glory, jest, and riddle of the world." To know ourselves is to know the information in the seed from which we sprouted -- we are at least in part the chemical expression of the message encoded in our genes. I suggest that there is encoded in our genes also a linguistic message, a message with semantic content, ard that Man is now in a position to decipher it.

The idea of a linguistic message reposing in our genes will likely be scoffed at by academic linguists and biologists, despite the enlightened age in which we live. This article therefore introduces the subject to recreational linguists for their deserved exclusive benefit. The "game" is to decipher the linguistic message of our genes. Astute readers will recognize areas for cooperative research by biologists, cryptologists, logologists and molecular oligomaths.

One approach is to "read" the genes in their biologically translated form as proteins. Genes dictate the primary structure of the proteins that make up our bodies. Proteins thus speak for genes. The primary structure of a protein is a unique permutation of tandem molecular subunits belonging to a small set of subunit types. A protein's subunit permutation is a precise decipherment of a message encoded in a corresponding gene -- a different gene species for each protein species. As I show below, to read the message linguistically, each permutation of subunits can be mapped to a permutation of alphabetic letters like that occurring in a written sentence.

The Molecular Alphabet of Proteins

Protein comprises the bulk of the body's organic substance. Tens of thousands and possibly hundreds of thousands of structurally different protein species (types) exist in the human body, each with a specific function. Each species occurs as an enormous number (more than 10^{10}) of identical clones (tokens). Their mass and interconnection provide the body's structural matrix, including that of bone. Inside cells, they function dynamically, organizing and expediting chemical reactions that extract energy from nutrients needed for growth and development, for defense against biodegradation, and for coordinated movement and thought (except when alcohol is taken in excess). In the bloodstream, circulating proteins transmit information among neighboring and remote organs spreading gossip and humours.

The figure below depicts the protein's primary structure in typewriter keyboard symbols:

```
   H      H      H      H      H      H      H      H      H      H
   |      |      |      |      |      |      |      |      |      |
]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)
   |      |      |      |      |      |      |      |      |      |
   R4     R9     R16    R16    R2     R7     R4     R9     R3     R9
```

A protein is like a string of vari-sized beads, viz., a string of vari-sized molecules called amino acids. Each amino acid molecule has a hub, depicted as *, attached to which (by -- or |) are four different chemicals, depicted as],), H, Rn. Only the Rn's differ in the repeating structure. Two components,] and), couple the hub to amino acids at either side, thus forming a linear chain. (The connections are called "peptide bonds", and the chain is called a "polypeptide" or "protein".) A third component, the so-called R-group, Rn (n=1,2,3,...20), lies at right angles to the chain and is not directly involved in linking. The fourth component, a single hydrogen atom, H, also is not involved in linking. The *,],) and H components do not differ along the chain. Although the R-groups do differ along the chain, any one or all may be used many times in the chain.

The R-groups distinguish the different amino acid types. There are twenty different amino acid types, each with a unique Rn (n=1, 2,3,...20). Each protein species has its own unique sequence of Rn's.

As Lehninger [1970] states in his textbook on biochemistry, "The R-groups [Rn's] are the 'letters' in the molecular alphabet of protein structure." Had Lehninger pursued the thought, he might have wondered whether those "molecular letters" have alphabetic equivalents (e.g., R1=S, R2=E, ...) that create words and sentences that reveal a linguistic message (see figure below).

```
   H      H      H      H      H      H      H      H      H      H
   |      |      |      |      |      |      |      |      |      |
]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)]--*--)
   |      |      |      |      |      |      |      |      |      |
   R4     R9     R16    R16    R2     R7     R4     R9     R3     R9

   H      E      L      L      O      T      H      E      R      E
```

Thus, the chemical structure of a protein is like the orthographic structure of a sentence. A sentence is a variably long string of letters: a permutation of selected tokens of unique letter-types (alphabetic letters: a,b,c, ...) drawn from a finite set (twenty-six letters). Every legitimate unique permutation defines a sentence-type (e.g., "Time flies like an arrow" or "Fruit flies like a banana"). Similarly, a protein is a variably long string of molecules ("amino acids"), consisting of a permutation of selected tokens (generally 50 to 1000 tokens) drawn from a finite set (twenty amino acid types: glycine, alanine, valine, ...). Every unique permutation defines a protein-type or species (e.g., insulin, hemoglobin, hobgobulin).

The analogy between sentences and proteins extends beyond the orthographic. As with sentence types, there is no theoretical limit to the number of protein types. (For a typical chain length of 100 amino acids there are 20^{100} (ca. 10^{130}) possible protein types.) The chain lengths of 50 to 1000 that are commonly encountered in human proteins correspond to the letter lengths of typical English sentences. In the same way that the infinite variety of sentence types enable diversity and specificity in language expression, the astronomically large (theoretically infinite) number of possible protein types enable diversity and specificity in biological expression.

Deciphering Requirements

Assuming that the linguistic genetic message is targeted to a written language that has an alphabet-based writing system, one must know three things to decipher the message: (1) the amino acid sequence (primary structure) of the protein, which defines the complete sequence of R-groups of the protein (e.g., R16-R11-R2-R7-R7-R9-R11...); (2) the R-group-to-letter code, indicating which R-group (or sequence of R-groups) corresponds to which alphabetic letter (e.g., R6=F, or R9-R3-R14=T); and (3) the target language (e.g., English, Sanskrit, etc.).

The biological activity of the amino acid sequences might provide a clue to the R-group-to-letter code. For most of the sequenced proteins, one or more biological functions are known; designation of biological function might be a part of the message (a kind of internal Rosetta stone). For example, for insulin, there may be a sequence saying "lowers blood sugar" or "facilitates glucose transport". The biologist's elucidation of a protein's biological function then may become part of the deciphering process.

But Speech is More Basic

Traditional linguists will have little patience with any exercise of "alphabetizing" the genetic code. They assert that alphabetic writing derived from speech as a symbolic record of speech; speech, the true foundation of language, antedates writing. The alphabet has too few elements (letters) to represent the full repertoire of human speech sounds (phonemes), they will point out. The genetic message is to speech, not writing, they will scream.

Yet their arguments cannot exclude written language as the primary target of the linguistic genetic message. According to Fromkin and Rodback [1988], "The Talmudic scholar Rabbi Akiba believed that the alphabet existed before humans were created; and according to Islamic teaching, the alphabet was created by Allah himself, who presented it to humans but not to the angels." Were those thinkers unconsciously "reading" their genetic message?

Speech can be viewed as rudimentary writing, neurologically simpler than handwriting, and therefore developing earlier in response to subconscious "reading" of our linguistic genetic message. You may laugh out loud at this, which only goes to prove the point.

Expressed another ways, the inherent drive of our genes to "write" themselves out with muscle power may emerge first as speech owing to the ease of vocalization. Babies babble. Developing fine motor and intellectual coordination for handwriting takes longer. How much longer it will take for deliberate "reading" of our genetic message depends on how soon these concepts are accepted.

Speechless humans might still have developed alphabetic writing. That idea may render academic linguists speechless themselves, but that's okay, since most of them write fairly well. Once writing has been learned, speech is dispensable, despite its richer repertoire of phonemes compared to letters. For a practiced reader, the words on this page need not suggest sounds at all. Thus, speech may have been our genetic intelligence's preliminary medium for expressing messages better expressed in a universal written language.

Who wrote the message? Perhaps it was a literate extraterrestrial who bioengineered the message into a seed that Earth could germinate. That sounds like science fiction but is consistent with the hypothesis of Directed Panspermia seriously proposed by the Nobel Prize winning scientist, Francis Crick. (Crick is no crock; he co-discovered the architecture of the gene, a nut not easily cracked.) Perhaps those fantastically advanced extraterrestrial bioengineers left their signature in the message, something like "Made in Galaxy ..." or "Made You Look!" stamped in a protein on the gluteus maximus. To think, we may have been sitting on the answer all along!

The Language of the Protein's Linguistic Message

In considering which language the linguistic genetic message might be written in, at least four possibilities emerge:

(a) A Novel Language. Like invented languages (Interlingua, Esperanto, Advertising, etc.), the target language might never have been native. It may have lain dormant in our genes for millennia, awaiting discovery.
(b) English. As the world's most universal and polyglot language, English is an obvious candidate.
(c) The Human Proto-Language. All of the world's languages may have derived from a single ancestor, an obvious candidate.
(d) A Mixture of Languages. Typical protein chain lengths are sufficiently long to contain the same message repeated in different languages.

If the linguistic genetic message is written in a novel language, deciphering it would perhaps be the most momentous event in the history of writing since God wrote on the wall with her finger.

The R-Group-to-Letter Code

Taking the case in which the message has been written in English, the problem of determining the R-group-to-letter code becomes one of mapping twenty amino acid types onto the twenty-seven character set of twenty-six letters (A,B,C,...Z) and the space character that separates words. Since the type-counts differ, a simple one-to-one mapping is not possible. The disparity can be reduced in several ways, and eliminated by a combination of them:

(a) The R-group set might contain no space character equivalent. Spacesbetweenwordsfacilitatereadingbutarerarelyneededtoeliminate-ambiguity. Spaces might be absent or might be coded by a unique sequence of R-groups never used in words (e.g., an unused trigram like QZQ).

(b) An alphabetic letter (e.g., "j" as in "jest") unambiguously representable by a pair of other letters (e.g., "dg" in "dgest") might have no representation in the R-group code. Thus, j, q, and x might not be R-group coded since j=d+g, q=k+w (kwick for quick), and x=k+s (aks for ax). "Kwickly swung the aks of dgustice." The letter "c" also might not be coded for, since either "k" or "s" can represent the c-sound (kar for car, sentury for century).

(c) Similarly, other instances of alphabetic redundancy might not be encoded. For example, "i" is redundantly represented in "y" and "w" is a redundancy in "u"'s: "UUolf bytes man."

Eliminating spaces and j, q, x, c, i, and w reduces the English alphabet to twenty characters, permitting a one-to-one mapping of R-groups and letters.

Another approach is to eliminate vowels: nthr pprch s t lmnt vwls. The earliest alphabetic writing systems did not represent vowel sounds. Having one of the twenty amino acids mark vowel positions makes the message easier to discern: th*s m*ss*g* w*s wr*tt*n w*th**t v*w*ls.

Chemists have assigned a unique one-letter abbreviation for each amino acid (to facilitate sequence-comparison among proteins). Perhaps they have unwittingly guessed the cipher. The chemist's one-letter code is shown below along with the chemical names and abbreviated chemical names of the amino acids.

Perhaps the relative frequency of occurrence of amino acids among protein types given in Klapper [1977] corresponds to the relative frequency of letters among sentence types given in Haldane [1976]. Thus, alanine, the most frequently used amino acid, may correspond to "e", the most frequently used letter, and so on down the list. Eliminating j, q, x, c, i and w, as discussed above, those frequency correspondences are shown in the last column of the table.

There are many other possibilities. Detailed information available about amino acid geometries might reveal a shape-to-shape cipher. Likewise, the molecular weights of the amino acids are known; the ascending-order weight-sorted amino acids might correspond to the similarly weight-sorted letters, in which letter-weights are assigned according to alphabetical position (e.g., A=0, B=1, C=2, ...Z=25).

Exercises for the Reader

Adventurous readers are encouraged to try their hand at deciphering the linguistic message in human proteins with known amino

AMINO ACIDS IN HUMAN PROTEINS				
% Frequency of Occurrence	Chemical Name	Abbreviated Chemical Name	One-letter Chemists' Code	Frequency-matched English Letter
9.0	Alanine	Ala	A	E
7.5	Leucine	Leu	L	T
7.5	Glycine	Gly	G	A
7.1	Serine	Ser	S	O
7.0	Lysine	Lys	K	N
6.9	Valine	Val	V	R
6.2	Glutamic acid	Glu	E	I or Y
6.0	Threonine	Thr	T	S or C
5.5	Aspartic acid	Asp	D	H
4.7	Arginine	Arg	R	D
4.6	Isoleucine	Ile	I	L
4.6	Proline	Pro	P	F
4.4	Asparagine	Asn	N	M
3.9	Glutamine	Gln	Q	U
3.5	Phenylalanine	Phe	F	G
3.5	Tyrosine	Tyr	Y	P
2.8	Cysteine	Cys	C	B
2.1	Histidine	His	H	V
1.7	Methionine	Met	M	K or C
1.1	Tryptophan	Trp	W	Z

acid sequences. For starters, listed below are the amino acid sequences of two human proteins. The first is that of alpha hemoglobin, one of two proteins in red blood cells that supplies oxygen to cells throughout the body. The second is that of stathmin, a ubiquitous intracellular protein involved in fundamental regulation of cellular viability. Additional sequences can be found in print in the atlas by Dayoff [1972] and in computer-readable form in several molecular biology databases.

Alpha hemoglobin (using abbreviated chemical names):
Val-Leu-Ser-Pro-Ala-Asp-Lys-Thr-Asn-Val-Lys-Ala-Ala-Trp-Gly-Lys-
Val-Gly-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met-
Phe-Leu-Ser-Phe-Pro-Thr-Thr-Lys-Thr-Tyr-Phe-Pro-His-Phe-Asp-Leu-
Ser-His-Gly-Ser-Ala-Gln-Val-Lys-Gly-His-Gly-Lys-Lys-Val-Ala-Asp-
Ala-Leu-Thr-Asn-Ala-Val-Ala-His-Val-Asp-Asp-Met-Pro-Asn-Ala-Leu-
Ser-Ala-Leu-Ser-Asp-Leu-His-Ala-His-Lys-Leu-Arg-Val-Asp-Pro-Val-
Asn-Phe-Lys-Leu-Leu-Ser-His-Cys-Leu-Leu-Val-Thr-Leu-Ala-Ala-His-
Leu-Pro-Ala-Glu-Phe-Thr-Pro-Ala-Val-His-Ala-Ser-Leu-Asp-Lys-Phe-
Leu-Ala-Ser-Val-Ser-Thr-Val-Leu-Thr-Ser-Lys-Tyr-Arg

Stathmin (using the one-letter chemist's code):
MASSD1QVKELEKRASGQAFEL1LSPRSKESVPEFPLSPPKKKDLSLEEIQKKLE-
AAEERRKSHEAEVLKQLAEKREHEKEVLQKAIEENNNFSKMAEEKLTHKMEANKE-
NREAQMAAKLERLREKDKHIEEVRKNKESKDPADETEAD

Direct Translation of the Gene

In using a protein's amino acid sequence to discover the linguistic message of its corresponding gene, we work with a translated message, gene-to-protein, not with the primary message itself. Something may get lost in translation, however, since not all the genetic material codes for protein. That could be embarrassing, depending on the message. We might be able to recover the whole message if we mapped the molecular elements of the gene directly to the orthographic elements of the language (letters, spaces, punctuation, numerals, etc.). That is left as an exercise for the reader.

BIBLIOGRAPHY

Dayoff, M.O. (ed.), 1972. Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington DC

Fromkin, V. and Rodman, R., 1988. An Introduction to Language, Fourth Edition, Holt, Rinehart and Winston, NY (p. 339)

Haldane, R.A., 1976. The Hidden World, St. Martin's Press, NY (pp. 149-155)

Klapper, M.H., 1977. Biochem Biophys Res Commun 78 (pp. 1018-1024)

Lehninger, A.L., 1970. Biochemistry: The Molecular Basis of Cell Structure and Function, Worth Publishers, NY (p. 67)

THE OMNI GAZETTEER

Well-heeled logologists will be delighted to hear that one can now obtain an eleven-volume listing of more than 1.5 million US placenames (not all different), compiled by Omnigraphics Inc. (Penobscot Building, Detroit MI 48226), based on the Geographic Names Information System plus other government databases. The price, alas, is $2000, in either the printed or CD-ROM version (available spring of 1992). For this, one gets not only the name, but also its latitude and longitude, the name of the USGS 7.5' series map on which it appears, its county, and the type of feature (city or town, school, park, shopping center, airport, body of water, cemetery, hill or mountain, historic place, church, etc.). Those interested in the names only can save money by buying the one-volume National Index for only $250. The detail is stupefying. I note under ACACIA California CANAL, DRAIN, FIVE A DRAIN, FIVE B DRAIN, FIVE DRAIN, LATERAL EIGHT, LATERAL FIVE, LATERAL FIVE A, and six more of similar ilk. Yet experts believe that there are more than 3.5 million current placenames in the US, as well as 3.0 million street names!