Protein J (2012) 31:550–563
DOI 10.1007/s10930-012-9432-7

Information Properties of Naturally-Occurring Proteins: Fourier Analysis and Complexity Phase Plots

Daniel J. Graham · Shelby Grzetic · Donald May · John Zumpf

Published online: 20 July 2012

Abstract In previous work from this lab, the information in natural proteins was investigated with Ribonuclease A (RNase A) serving as the source. The signature traits were investigated at three structure levels: primary through tertiary. The present paper travels further by charting the primary structure information of about half a million molecules. This was feasible given abundant sequence archives for both living and viral systems. Notably, a method is presented for evaluating primary structure information, based on Fourier analysis and spectral complexity. Significantly, the results show certain complexity traits to be universal for living sources. Viruses, by contrast, encode protein collections which are case-specific and complexity-divergent. The results have ramifications for discriminating collections on the basis of sequence information. This discrimination offers new strategies for selecting drug targets.

Keywords Proteins · Fourier analysis · Spectral entropy · Complexity · Phase diagrams · Viruses

Abbreviations

PDB	Protein data bank
RNase A	Ribonuclease A
QSAR	Quantitative structure activity relation
CI	Correlated information
WT	Wild type
2D	Two dimensional

1 Introduction

Amino acid sequences are the blueprints for proteins. Their information is readily quantified in high-level terms. If a sequence is restricted to the 20 standard amino acids (**G**, **V**, **A**, …), and unrestricted diversity-wise, each constituent is allied with $\log_2(20) \approx 4.32$ bits of information. A decapeptide requires 43.2 bits to specify; a 100-unit molecule calls for 432 bits. Proteins built from hundreds of amino acids are more complicated than small systems (tens of amino acids) because, for openers, their sequences hold more information.

Yet the issues run deeper when a sequence is tied to a chemical function, say, glycoside hydrolysis. If the catalytic action necessitates isoleucine (**I**) in the 23rd position (in boldface, counting from left):

KVFERCELARTLKRLGMDGYRG**I**SLANWMCLAK
WESGYNTRATNYNAGDRSTDYGIFQINSRYWCN
DGKTPGAVNACHLSCSALLQDNIADAVACAKRV
VRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV
(Seq. 1)

it is only because of details in the preceding 22- and succeeding 107-member strings, namely:

KVFERCELARTLKRLGMDGYRG (Seq. 2)
SLANWMCLAKWESGYNTRATNYNAGDRSTDYG
IFQINSRYWCNDGKTPGAVNACHLSCSALLQDNI
ADAVACAKRVVRDPQGIRAWVAWRNRCQNRDV
RQYVQGCGV (Seq. 3)

Put another way, the *need* for **I** in the 23rd position, and not **G**, **A**, **D**, etc. hinges on *all* the other constituents. In effect, every member of Seq. 1 depends on 129 others; each is impacted by the information of neighbors near and far. To be sure, structures such as Seq. 1—which describe

D. J. Graham (✉) · S. Grzetic · D. May · J. Zumpf
Department of Chemistry, Loyola University Chicago,
6525 North Sheridan Road, Chicago, IL 60626, USA
e-mail: dgraha1@luc.edu

lysozyme—are nonsensical in appearance [21]. Yet if a protein researcher understood the assembly rationale, he or she could infer portions of the sequence upon learning others. Such comprehension would *lower* the bits needed for specifying 130 letters.

Yet the information issues run deeper still. The workings of lysozyme do not transpire in a vacuum, but rather in an environment of working proteins. The following are the respective sequences for adenylosuccinate synthetase, triphosphate epimerase, and Ribonuclease A (RNase A) [13, 23, 25]:

MSGTRASNDRPPGTGGVKRGRLQQEAAATGSRV
TVVLGAQWGDEGKGKVVDLLATDADIVSRCQGG
NNAGHTVVVDGKEYDFHLLPSGIINTKAVSFIGNG
VVIHLPGLFEEAEKNEKKGLKDWEKRLIISDRAHL
VFDFHQAVDGLQEVQRQAQEGKNIGTTKKGIGPT
YSSKAARTGLRICDLLSDFDEFSARFKNLAHQHQS
MFPTLEIDVEGQLKRLKGFAERIRPMVRDGVYFM
YEALHGPPKKVLVEGANAALLDIDFGTYPFVTSSN
CTVGGVCTGLGIPPQNIGDVYGVVKAYTTRVGIG
AFPTEQINEIGDLLQNRGHEWGVTTGRKRRCGWL
DLMILRYAHMVNGFTALALTKLDILDVLSEIKVGI
SYKLNGKRIPYFPANQEILQKVEVEYETLPGWKA
DTTGARKWEDLPPQAQSYVRFVENHMGVAVKW
VGVGKSRESMIQLF (Seq. 4)
MAQPAAIIRIKNLRLRTFIGIKEEEINNRQDIVINVTI
HYPADKARTSEDINDALNYRTVTKNIIQHVENNRF
SLLEKLTQDVLDIAREHHWVTYAEVEIDKLHALR
YADSVSMTLSWQR (Seq. 5)
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSR
NLTKDRCKPVNTFVHESLADVQAVCSQKNVACK
NGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTT
QANKHIIVACEGNPYVPVHFDASV (Seq. 6)

By contrast, the following describe molecules which have been constructed arbitrarily by the authors and thus pose no known function:

MHACYRRDTWHFPQSRRHARSEWDLEDVGTHHY
QGCEPKERNFWSEYKKQKQKGQGFALHQYERIQI
KPAWSNIMLDHASINGESDFHHDILTMHFDERCQ
QTFPYHYVGQPPNPWILLRNETCDVTTKEPNFTTD
RQQHFGHNGMHFEQMKDHKFMHQFESEIEDCTF
WWLPWKWPNWYGKPFNHEWPMYAACWIINDH
CIMVRWYDHFRTD (Seq. 7)
MCMNRIIEIREGKPNEMVWPEDCHCRVPRQYECC
QYVQFMVWFLYNGFMDFCGVHYDTGGEDLWLN
HLQIKYPPSESHQTGCVYVCLMVPRFRTESVHAC
HVMWFYWKCDHPLCCQRDGHVRCVWSLAAGQL
RHFPRYQTLTSFPQQLLPPISQIIFISQYVVIEWKIQP
CVFEQMWISVDGADRYQDKLPGYHRREQNKSSD
ELLVMVEYRLKLCVNVI (Seq. 8)
MTMYHPCWKSDRHKQALMNKGPWYPDEKTEIYF
RHLWPITWPGVAFPWEQNKSDGHPTQHKRELGG
YNHNNFWGLGDYIGPVKSVAEDWVFWQCEEMK
QMMAKIAKLVKFGVGWYDQFFTCGDVNWHFVIG
WLYTHMCQIDNNEGGLHHACAASTMPTAFYFAD
DKGFGGIRVPDMPPNPWNSEQVEMKTEWAAKKD
IWDPFEGFWGGHNYYMTEYNIKPPEVPWMPNQA
VVEVKRKMTCCYIGCLGVKYAVDDYCTPWSVQG
KIQN (Seq. 9)

Countless examples can be presented. The point is that the design rationale for lysozyme applies to (naturally occurring) Seqs. 4–6 as well. For not only does an amino acid impact others in a molecule, there are intermolecular relationships as well. If a researcher understood these, he or she could infer parts of lysozyme upon learning Seqs. 4–6. An investigator would surmise as well that {Seqs. 1, 4, 5, 6} and {Seqs. 7, 8, 9} really constitute two distinct sets: for naturally-occurring and arbitrarily-assembled molecules, respectively.

At present, databases store the primary structures for $\sim 10^7$ proteins. There appears no limit given the diversity of molecules and their species of origin. If the average protein features 300 amino acids, then databases provide sequence information aplenty:

$$10^7 \times 300 \times 4.32 \approx 1.3 \times 10^{10} \text{ bits.}$$

At the same time, the number of 300-unit possibilities is astronomical at $20^{300} \approx 10^{390}$. Clearly, only nature's surface has been scratched and the same statement will apply when databases have grown ten-fold. At the least, a better understanding of protein assembly rationales is desirable because it can save significant labor in the future.

The messages in amino acid chains are the object of international research. Efforts cut intensely across chemistry, biology, mathematics, and medicine. For our part, we have been investigating archetypal systems with the help of information theory. Thus far, our reports have looked to RNase A (Seq. 6) for insights [9, 10]. Ref. [9] focused on the information scaling of the RNase A primary structure, in comparison with sequence isomers. The appendix of that work illustrated how similar scaling attributes were expressed by proteins carrying radically different chemical functions. This was important. It showed that the information distribution of a *particular* catalyst (RNase A) with a dedicated and evolved chemical function was not merely case-specific.

Reference [10] delved more deeply by examining information in the RNase A secondary and tertiary structure. As is well known, these structure levels are subsidiary to the primary and confer the molecule's chemical action. The results identified the information specialness of the wild-type (**WT**) system, compared with mutants.

The present paper takes the next steps as the information properties of multiple systems (i.e. not just RNase A) are examined. Further, the information specialness of **WT** (i.e. naturally occurring) systems is probed more deeply. A variety of living organisms and viruses furnish the systems of interest.

Note the crux issue. On one hand, it is easy to construct arbitrary (non-functional) sequences such as Seqs. 7–9: one can write letter strings at whim or assemble them using a random number generator. For the motivated researcher, there is a real molecule waiting to be synthesized and studied on the basis of each. Further, nature is generous with new and raw material for sequencing labs. As a result, there are ever-growing structure archives.

On the other hand, it is challenging to discriminate natural and arbitrary sequences, e.g. Seqs. 4 versus 7. Appeal can always be made to databases and search engines. Yet pinpointing Seq. 4 as a synthetase in a BLAST search, while discovering Seq. 7 not to align with database entries does not further our understanding of design rationales. For example, how can one discriminate which of the following is the functional sequence?

MKNKYNVLCRWMKLCNWKMGKHRWGYPRCHMVTKMADRISGCNEVKWVEPECELRLCNRTHLQRWGRLYGCFVIFQILHISWFHIVNVTSTGNQKYKHNEQMVRYFCEKKMFSNLGNRHGSCCGFEQGMAAHDNNCPSRHDLGCRKEQHQETTSMMSSPCKPMWSGPYVPEKYRVSAGNYNGVCRQMHKQPSCHLNMYNFFLCMAWIIREPKSLYKQKHALCNWNRQEPFKEANYICMNRDYQWV (Seq. 10)

MNPESRVIRKVLALQNDEKIFSGERRVLIAFSGGVDSVVLTDVLLKLKNYFSLKEVALAHFNHMLRESAERDEEFCKEFAKERNMKIFVGKEDVRAFAKENRMSLEEAGRFLRYKFLKEILESEGFDCIATAHHLNDLLETSLLFFTRGTGLDGLIGFLPKEEVIRRPLYYVKRSEIEEYAKFKGLRWVEDETNYEVSIPRNRIRHRVIPELKRINENLEDTFLKMVKVLRAEREFLEEEAQKLYKEVKKGNCLDVKKLKEKPLALQRRVIRKFIGEKDYEKVELVRSLLEKGGEVNLGKGKVLKRKERWLCFSPEV (Seq. 11)

The answer appears in this paper wherein a rapid and effective discrimination method is presented. Suffice to say that functional proteins are special by their recognition and catalytic capabilities. The protein collection of a cell or virus is extra-special by its capacity to generate high-fidelity copies. Clearly the strategies for linking amino acids are anything but arbitrary.

The present paper looks beyond RNase A and is based on the information properties of half a million compounds. The calculations are extensions of ones introduced in Ref. [9]. The results are new insights into proteins as grounded upon Fourier spectra and two-dimensional (2D) phase plots. Such tools are timely, having close relatives developed by multiple research groups. In early 2011, Randić et al. [26] contributed an extensive review of the graphical and spectral representations of natural proteins. These researchers aimed at providing "novel mathematical and structural invariants that can serve as additional mathematical descriptors for such systems." Along complementary lines, González-Díaz et al. [1, 5, 7, 8] established the Markov and entropic fingerprints attendant to kinases and ribonucleases. Looking substantially beyond proteins, González-Díaz et al. have developed Markov and Shannon entropy models for connectivity within complex networks which *include* drug-target, host-parasite, cerebral cortex, and legal-social [24, 27]. Most recently Glisic and co-workers have applied Fourier spectrum and information methods in the assessment of viral protein sequences [6]. Bajorath and co-workers have applied 2D similarity methods toward pinpointing structure selectivity and molecular fingerprints [31]. Their research continues a multi-year effort to discriminate natural from unnatural libraries [2, 29]. We call attention also to the phylogenetic analysis of Chang and Wang. Their methods established information insights for diverse sources of proteins [4]. Likewise, Liu and co-workers leveraged Fourier and spectral techniques toward the prediction of membrane protein types [18]. In a similar vein, Jiang and co-researchers established prediction methods grounded on wavelet transforms [14]. Fourier and related analytical tools support central themes of the present paper.

2 Materials and Methods

Proteins are chains of amino acids with formulae encapsulated by letter sequences. The amino acid and letter correspondence is 1:1 with the left- and right-most entries representing N- to C-terminal units, respectively. Taken at face value, each member of a sequence offers the same amount of information.

Tying a sequence to the Angstrom level brings additional considerations into play, however. This is because different letters, say **W** and **A**, do not signify electronic assemblies of equivalent size and complexity. A protein is threaded by a backbone of α-carbon, carbonyl, and amide linkages. The electronics are tuned via the side-chain attachments to the α-carbons. Tryptophan (**W**) and glutamine (**Q**) host substantive attachments via 6:5 fused aromatic rings and [$NH_2CO(CH_2)_2$], respectively. The tuning is modest at alanine (**A**) sites by way of a single methyl group. Glycine (**G**) sites are the sparsest as the α-carbons feature only hydrogen atoms.

Electronics and information are intertwined. Thus, when a sequence is connected to the Angstrom level, the information needs to reflect more significant upticks where **W**s and **Q**s appear, compared with **A**s and **G**s. In effect, sequences such as for lysozyme, adenylosuccinate synthetase, etc. delineate side-chain tuning which modulates the information in such a way that confers the molecular activity.

The groundwork for this paper and its predecessors was laid by previous research in this lab, wherein the correlated information *CI* for naturally occurring amino acids was quantified [11]. Such was based on the atom and covalent bond structure of each entity and a Brownian computation model. An average value <*CI*> and standard deviation σ_{CI} were established for a library constructed around the twenty standard amino acids. A dimensionless quantity $Z_{CI}^{(i)}$ was formulated for each library member based on its *CI* relative to the average, viz.

$$Z_{CI}^{(i)} = \frac{CI^{(i)} - \langle CI \rangle}{\sigma_{CI}} \tag{1}$$

The (*i*)-superscripts in Eq. (1) refer to the various library members: **G**, **A**, **I**, **D**, … The results in Ref. [11] establish dimensionless $Z_{CI}^{(i)}$ for every entry of a sequence:

$$\begin{array}{llll} Z_{CI}^{(W)} = +2.63 & Z_{CI}^{(Y)} = +1.51 & Z_{CI}^{(H)} = +1.12 & Z_{CI}^{(C)} = +1.10 \\ Z_{CI}^{(F)} = +0.691 & Z_{CI}^{(N)} = +0.405 & Z_{CI}^{(D)} = +0.308 & Z_{CI}^{(Q)} = +0.168 \\ Z_{CI}^{(E)} = +0.115 & Z_{CI}^{(M)} = -0.128 & Z_{CI}^{(T)} = -0.286 & Z_{CI}^{(A)} = -0.476 \\ Z_{CI}^{(R)} = -0.565 & Z_{CI}^{(S)} = -0.618 & Z_{CI}^{(K)} = -0.687 & Z_{CI}^{(V)} = -0.855 \\ Z_{CI}^{(P)} = -1.01 & Z_{CI}^{(L)} = -1.15 & Z_{CI}^{(I)} = -1.43 & Z_{CI}^{(G)} = -1.49 \end{array} \tag{2}$$

The $Z_{CI}^{(i)}$ have nothing to do with the population statistics of amino acids. Rather, they have everything to do with the atom and covalent bond assemblies, and the correlations effected by α-carbon attachments. The signs reflect whether the information impact is above or below nature's library average. The magnitudes are established in units of the library standard deviation per Eq. (1). For example, **W** (tryptophan) offers a nearly $+3\sigma$ information impact in a protein while the effects of **G** (glycine) are substantially smaller, almost 1.5σ below average. The above $Z_{CI}^{(i)}$ values were applied extensively to Ribonuclease A in Ref. [9] and calculations for the present paper.

The information functions for proteins are straightforward. A dimensionless function $G(k)$ is constructed by adding $Z_{CI}^{(i)}$ terms in the order presented by a molecule; *k* is the counting index while the superscript labels apply to **G**, **A**, **V**, etc. The N-to-C terminal direction is of no more chemical significance than C-to-N. For consistency, however, we elect the former direction in all analyses. In so doing, $G(k)$ tracks the left-to-right *cumulative* information of a protein, taking the Angstrom-scale correlations into account:

$$G(k) = Z_{CI,1}^{(i)} + Z_{CI,2}^{(i)} + Z_{CI,3}^{(i)} + \cdots + Z_{CI,k}^{(i)} = \sum_{j=1}^{k \le N} Z_{CI,j}^{(i)} \tag{3}$$

$G(k)$ is composition- and sequence-dependent. Its value at integer variable *k* depends on *k*-number of amino acid units registered left-to-right. There is a unique $G(k)$ for every possible sequence.

The upper panel of Fig. 1 illustrates the workings of $G(k)$, based on the synthetase of Seq. 4. One is struck not only by the linearity of the function, but also that it trends significantly below zero. The former trait reflects the linear scaling of the cumulative information. The latter attests the predominance of low information amino acids: ones having $Z_{CI}^{(i)} < 0$.

Now *perfect* linearity of $G(k)$ (i.e. coefficient of determination $R^2 = 1$) would apply to a molecule built from a single type of unit, say, **A**. Fortunately, nature opts for complicated systems. For a real protein, the side-chain tuning of each α-carbon uniquely modulates $G(k)$. In turn, the function fluctuates about straight-line behavior.

Regression analysis leads to a second function $L(k)$: that for the best-fit line allied with $G(k)$, viz.

$$L(k) = mk + b \tag{4}$$

In Eq. (4), *m* and *b* are the respective slope and *y*-intercept of the regression line. The dotted line in the upper panel of Fig. 1 illustrates $L(k)$ for the synthetase (Seq. 4) with $m \approx -0.416$, $b \approx -8.20$, and $R^2 \approx 0.997$. $L(k)$

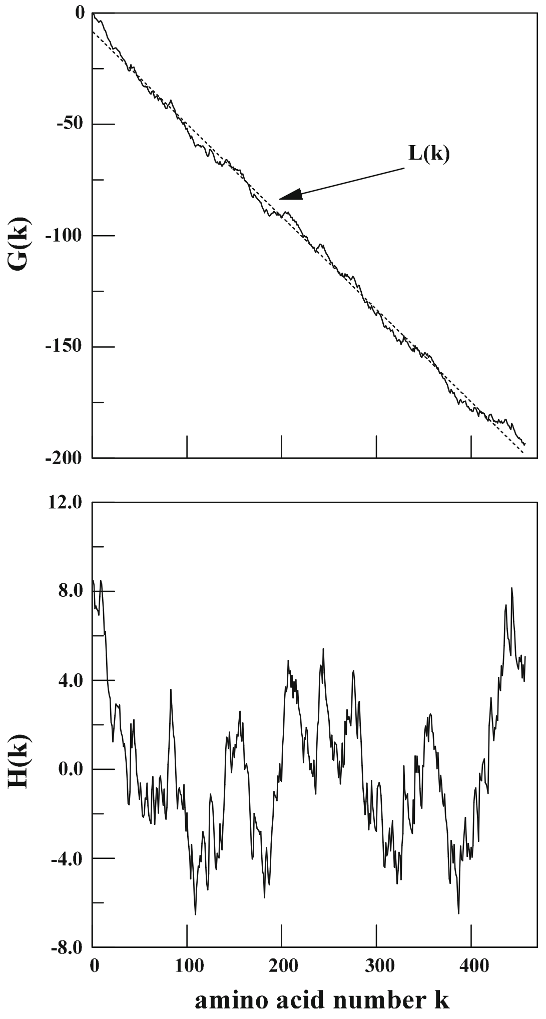


Fig. 1 $G(k)$, $L(k)$ (*dotted line*), and $H(k)$ for Sequence (4), adenylosuccinate synthetase. The quantities along the *horizontal* and *vertical scales* are dimensionless

enables the information modulations to be isolated via a third function $H(k)$:

$$H(k) = G(k) - L(k) \tag{5}$$

Like its parents $G(k)$ and $L(k)$, $H(k)$ is unique to a molecule. And importantly, $H(k)$ details the patterns in the information distribution. $H(k)$ equals zero for "single-letter" compounds such as polyalanine, polyglycine, and so on. As seen in Fig. 1, $H(k)$ traces out a mode structure of diverse frequencies and amplitudes.

Since $G(k)$ and $H(k)$ are evaluated at equidistant intervals, they accommodate Fourier transformations. To set the stage, the functions are shifted leftward by one unit:

$$\begin{aligned} \hat{G}(0) &= G(1) \\ \hat{G}(1) &= G(2) \\ &\vdots \\ \hat{G}(k-1) &= G(k) \end{aligned} \tag{6}$$

The results retain the same features as $G(k)$. The shift-operations apply just as easily to H so as to yield $\hat{H}$.

Every nth element of a *discrete* Fourier transform then arrives as follows:

$$\tilde{G}(n) = \sum_{k=0}^{N-1} \hat{G}(k) \exp\left[\frac{2\pi ikn}{N}\right] \tag{7}$$

$$\tilde{H}(n) = \sum_{k=0}^{N-1} \hat{H}(k) \exp\left[\frac{2\pi ikn}{N}\right] \tag{8}$$

In Eqs. (7) and (8), $i = \sqrt{-1}$ while π has its usual meaning [3].

$\tilde{G}$ and $\tilde{H}$ are unique to a molecule, just as their predecessors $G(k)$ and $H(k)$. At the same time, $\tilde{G}$ and $\tilde{H}$ are complex having both real and imaginary parts. Importantly, they re-formulate a protein's information in an alternate yet equivalent way. There is no loss of chemical message in taking $G \rightarrow \hat{G} \rightarrow \tilde{G}$ and $H \rightarrow \hat{H} \rightarrow \tilde{H}$. One can always backtrack so as to recover the amino acid sequence.

And there are other characteristics to note. In exercising the Fourier summations, the information of *every* constituent of a protein contributes to a point in $\tilde{G}(n)$ and $\tilde{H}(n)$. The synthetase responsible for Fig. 1 requires 457 amino acids for construction. Every one of the 457 impacts every part of $\tilde{G}$ and $\tilde{H}$.

For compression and ease of interpretation, the Fourier transforms are converted to power spectra. These arrive by computing the absolute values of $\tilde{G}$ and $\tilde{H}$ and taking symmetry into account [19]:

$$P_{\tilde{G}}(f_n) = \frac{1}{N^2} \cdot \left[\, \left|\tilde{G}(n)\right| \; + \; \left|\tilde{G}(N-n)\right| \right] \tag{9}$$

$$P_{\tilde{H}}(f_n) = \frac{1}{N^2} \cdot \left[\, \left|\tilde{H}(n)\right| \; + \; \left|\tilde{H}(N-n)\right| \right] \tag{10}$$

The frequency f_n is defined simply as:

$$f_n = \frac{n}{N} \tag{11}$$

Figure 2 illustrates the spectra descendent from Fig. 1. In effect, the power spectra distill the mode structures which are imbedded in the protein information. In the Fourier analysis of electric signals, frequency v and time t serve as conjugate variables [3, 19]. The variables f_n and k play analogous roles in the analyses at hand.

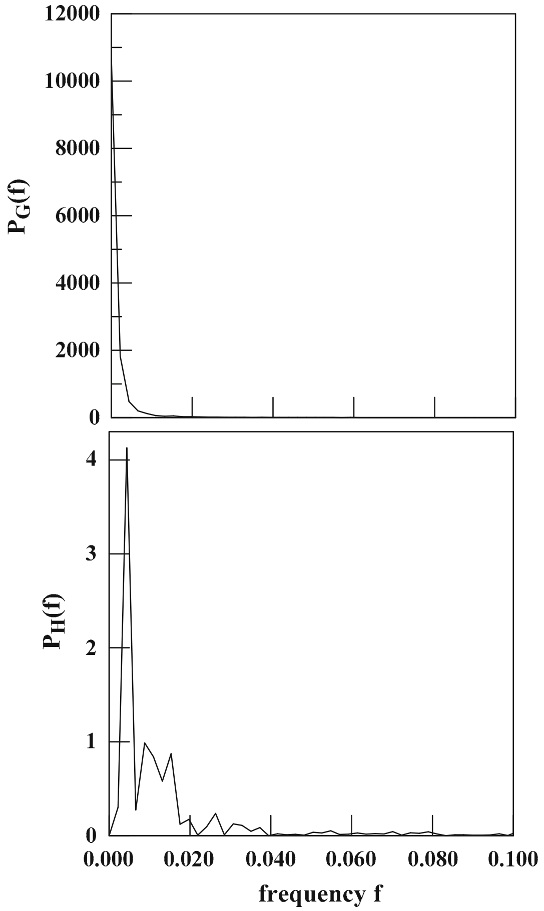


Fig. 2 Power spectra based on the synthetase data of Fig. 1

The Fig. 2 data are typical. They show the spectrum derived from $\tilde{G}(n)$ (upper panel) to be dominated by a single band with maximum intensity at $f_n = 0$. By comparison, the mode structures at higher frequencies prove only of minor intensity. These traits reflect (again) the linear scaling of $G(k)$ in Fig. 1.

The spectrum based on $\tilde{H}$ tells a longer story (Fig. 2, lower panel). Here a modulation structure of diverse frequencies and amplitudes is traced out. As is typical, the low-frequency modes contribute the majority of intensity. These bands tie especially to the long-range tuning of information in the molecule.

Every amino acid sequence, natural and contrived, offers plots as in Figs. 1 and 2. Thus the investigation of large populations (>5 × 10^5 for this paper) generates functions and spectra in true abundance. It then proves both convenient and enlightening to compress the data further. This obtains by computing spectral entropy values—measures of the Fourier image complexity. The power spectra are first subject to normalization:

$$\lambda \cdot \sum_n P_{\tilde{G}}(f_n) = 1 \qquad (12)$$

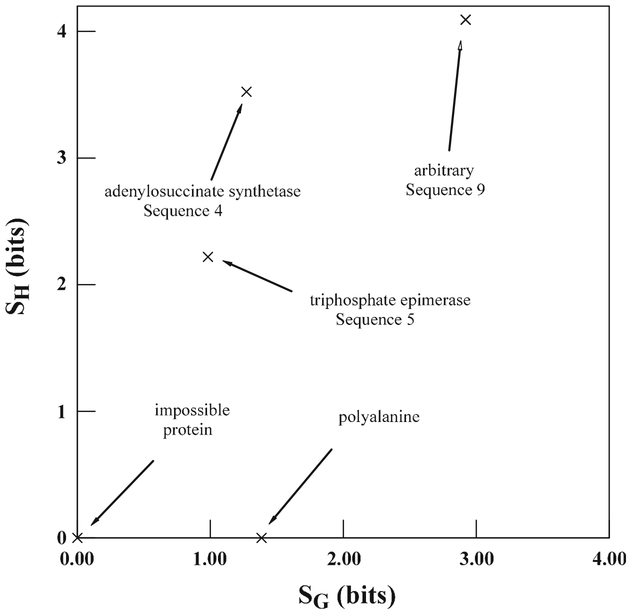


Fig. 3 State points located by information spectral entropy values

$$\mu \cdot \sum_n P_{\tilde{H}}(f_n) = 1 \qquad (13)$$

The parameters λ and μ are determined in each case by the summation details. Two entropy values, quantified in bits, arrive subsequently:

$$S_{\tilde{G}} = \sum_n \lambda P_{\tilde{G}} \cdot \log_2\left(\lambda P_{\tilde{G}}\right) \qquad (14)$$

$$S_{\tilde{H}} = \sum_n \mu P_{\tilde{H}} \cdot \log_2\left(\mu P_{\tilde{H}}\right) \qquad (15)$$

In effect, $S_{\tilde{G}}$ gauges the complexity of a protein's cumulative information while $S_{\tilde{H}}$ performs likewise for the modulation structure. Together, $S_{\tilde{G}}$ and $S_{\tilde{H}}$ locate a state point on a 2D phase plot with Fig. 3 illustrating sample results.

Figure 3 features several points on the $S_{\tilde{G}}$, $S_{\tilde{H}}$-plane. One point corresponds to the synthetase responsible for the preceding figures. Another state point is located by applying the same procedure to Seq. 5 detailing triphosphate epimerase. The data show $S_{\tilde{G}}$ and $S_{\tilde{H}}$ to fall in the range of one to a few bits. Amino acid sequences pose infinite possibilities. As a consequence, there are infinite possible pairings of $S_{\tilde{G}}$ and $S_{\tilde{H}}$.

Points for two of these infinite possibilities are included in Fig. 3. The state point for a polyalanine, $\mathbf{A}_{100}$, appears at zero altitude. The ground floor attests that homopolymers ($\mathbf{G}_{130}$, $\mathbf{W}_{225}$, etc.) all yield $P_{\tilde{H}}(f_n)$ with zero amplitude and zero $S_{\tilde{H}}$. Meanwhile, the state point for an impossible system falls at the origin. The forbidden-status reflects that no amino acid in nature's library evinces $Z_{CI}^{(i)} = 0$; therefore a protein with zero $G(k)$and $H(k)$ can never manifest. In contrast, arbitrarily-constructed molecules such as Seq. 9 yield state points *somewhere* on the $S_{\tilde{G}}$, $S_{\tilde{H}}$ plane. Moreover, every collection, natural and otherwise, defines a locus of points.

From the project outset, the authors were intensely curious about the point loci for natural versus arbitrarily-constructed proteins. What would they look like?

3 Results

Figure 4 illustrates the phase plot for the proteins manufactured by *Escherichia coli* (*E. coli*) cells. The sequences—as with all examined for this paper—were retrieved as FASTA files from the National Center for Biotechnology Information (NCBI) and subjected to the Fourier methodology of the previous section. The plot shows the placement of just under 5,000 state points, with the locus most concentrated near $S_{\tilde{G}}$, $S_{\tilde{H}} \approx 1.5, 3.0$ bits. The ranges prove as follows:

$$0.502 \le S_{\tilde{G}} \le 3.09 \text{ bits}$$

$$0.391 \le S_{\tilde{H}} \le 5.18 \text{ bits}$$

In effect, these establish complexity boundaries for this much-studied library of molecules. $S_{\tilde{G}}$, $S_{\tilde{H}}$ pairings outside the boundaries are certainly possible. But they are declined or avoided outright by the bacteria. Interestingly, the point locus has an arrowhead shape which is concentrated at the center while diffuse at the edges. A cleft structure moreover appears near $S_{\tilde{G}} \approx 1.40$ *bits* and $S_{\tilde{H}} \le 2$ *bits*. Note that the state points for arbitrarily-constructed molecules such as Seqs. 7–9 fall considerably outside the arrowhead territory. The $S_{\tilde{H}}$ values appear well within bounds, but the $S_{\tilde{G}}$ land too far to the right. The same statement applies to Seq. 10 of the Sect. 1 and affirms its unnatural status. By contrast, Seq. 11 details a lysidine synthetase—and thus naturally occurring molecule; this places a state point in the middle of arrowhead territory [22].

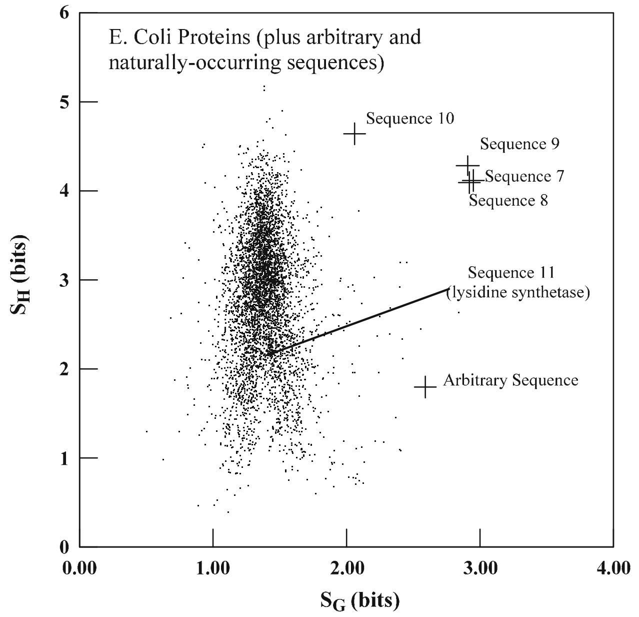


Fig. 4 Complexity phase plot for the *E. Coli* library. Included are the state points for other proteins such as presented in *text*. The coordinates are determined by spectral entropy values based on amino acid sequences. The points for arbitrary sequences lie outside the *E. Coli* point locus

Immediate questions come to mind. Are the complexity phase plots for other natural sources different from Fig. 4? Does a plot for one protein library match another only after rescaling $S_{\tilde{G}}$ and $S_{\tilde{H}}$? Quite a few inquiries motivated the authors to examine multiple sources.

Some answers appear in Figs. 5 and 6. The first is that the proteins for diverse *living* systems—algae, mice, humans, and fruit-flies—demonstrate highly similar $S_{\tilde{G}}$, $S_{\tilde{H}}$ plots—there is no need to rescale (i.e. compress or dilate) the coordinate axes. The same complexity territory is staked out while the arrowhead and cleft motifs are conserved across the boards. The differences lie in the dispersion of points, especially at the arrowhead edges. The authors investigated the protein libraries for twenty-three living systems. All yielded $S_{\tilde{G}}$, $S_{\tilde{H}}$ plots that closely mimicked one another as in Fig. 5.

Sharp contrasts were drawn by viruses, however. These systems encode multiple proteins, yet reside only at the threshold of life [17]. Figure 6 illustrates data for the libraries encoded by Influenza A, Human Immunodeficiency (HIV), Hepatitis B, and Simian (monkey) Influenza viruses. The arrowhead morphology is altered significantly in all four cases while multiple island structures appear. The latter owe to local concentrations of state points. The spectral entropy ranges veer substantially from living systems. For Influenza A, for example:

$$0.714 \le S_{\tilde{G}} \le 2.12 \text{ bits}$$

$$0.678 \le S_{\tilde{H}} \le 5.00 \text{ bits}$$

Clearly, viral proteins stake out an information territory with somewhat fluid boundaries. Even so, the territory barely encroaches on the state point region of arbitrarily-constructed molecules such as Seqs. 7–10.

Phase plots can be compared by visual inspection. Quantitative comparisons arrive by computing overlap values. Let two protein libraries express spectral complexity distributions Φ_A and Φ_B. Let $\Phi_A \cdot \Delta S_{\tilde{G}} \Delta S_{\tilde{H}}$ and $\Phi_B \cdot \Delta S_{\tilde{G}} \Delta S_{\tilde{H}}$ signify the fraction of library members with state points in a small window bounded along the horizontal by $S_{\tilde{G}}$ and $S_{\tilde{G}} + \Delta S_{\tilde{G}}$; let the vertical window boundaries be $S_{\tilde{H}}$

Fig. 5 Complexity phase plots for protein libraries as labeled

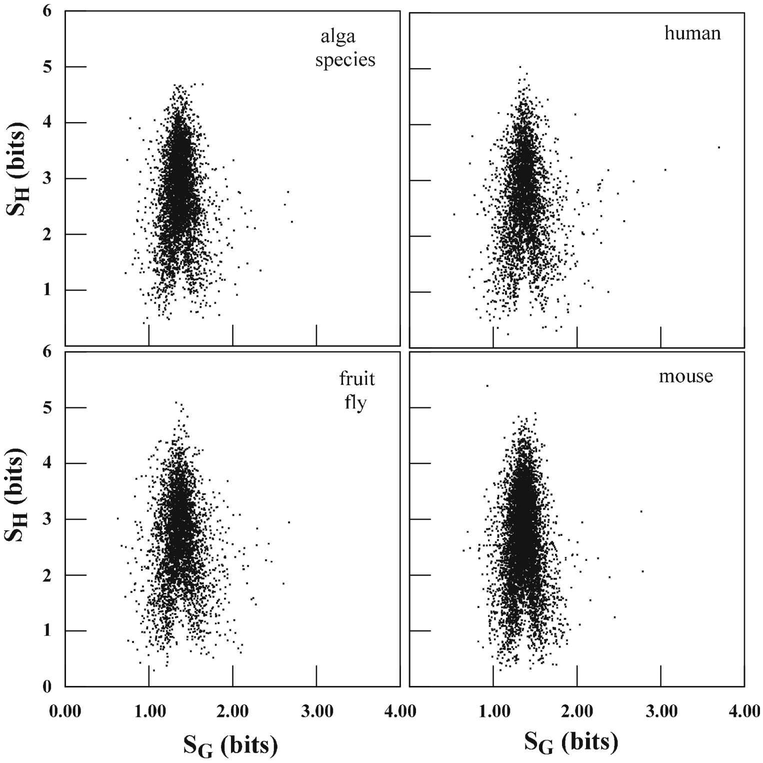
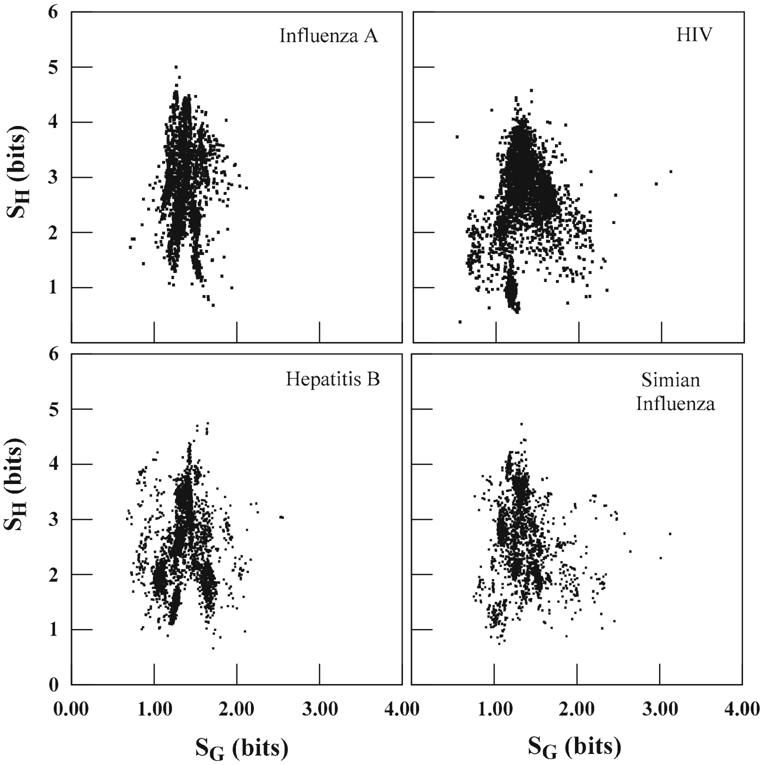


Fig. 6 Complexity phase plots for viral protein libraries



and $S_{\tilde H} + \Delta S_{\tilde H}$. Now if Φ_A and Φ_B were identical, their overlap would be the maximum possible. Thus, when comparing phase plots quantitatively, a dimensionless overlap O_{AB} needs to be computed as follows:

$$O_{AB} = \frac{\sum_{S_{\tilde G}}\sum_{S_{\tilde H}} \left(\Phi_A \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}\right)\cdot\left(\Phi_B \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}\right)}{\left[\sqrt{\sum_{S_{\tilde G}}\sum_{S_{\tilde H}} \left(\Phi_A \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}\right)^2}\right]\times\left[\sqrt{\sum_{S_{\tilde G}}\sum_{S_{\tilde H}} \left(\Phi_B \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}\right)^2}\right]} \tag{16}$$

Equation (16) has its roots in statistical structure investigations of liquid phase samples [12]. In directing Eq. (16) to proteins, one randomly selects an identical number of state points, say 3,500, for A and B libraries. The point fractions are then tallied for each window. The denominator in Eq. (16) ensures that O_{AB} equals 1 at maximum whilst the minimum possible is 0 for phase plots that diverge wholesale.

Figure 7 illustrates overlap results for multiple libraries. In assembling the data, the *E. coli* plot (Fig. 4) has been used as the reference: $\Phi_A = \Phi_{E.\ coli}$ while Φ_B is varied. For clarity, the data are presented as a left-to-right ascendant distribution. The horizontal error bars mark the average plus/minus one standard deviation. The errors are estimated by computing O_{AB} for several randomly-selected populations of each library.

The trends are striking. The overlap between living systems and *E. coli* is substantially greater compared with viral systems, e.g. $\left\langle O_{fruitfly,\ E.\ coli}\right\rangle = 0.6054 \pm 0.0043$whereas $\left\langle O_{Influenza\ A,\ E.coli}\right\rangle = 0.232 \pm 0.00443$. The phase plots for *ostreococcus-tauri,* an algae species, and *arabidopsis thaliana* (a common weed) demonstrate the greatest overlap with *E. coli* (bacteria). Regarding living sources, the protein libraries for cows and humans (mammals) demonstrate the least overlap with *E. coli*. The overlap differences among living systems prove small but (in most cases) statistically significant.

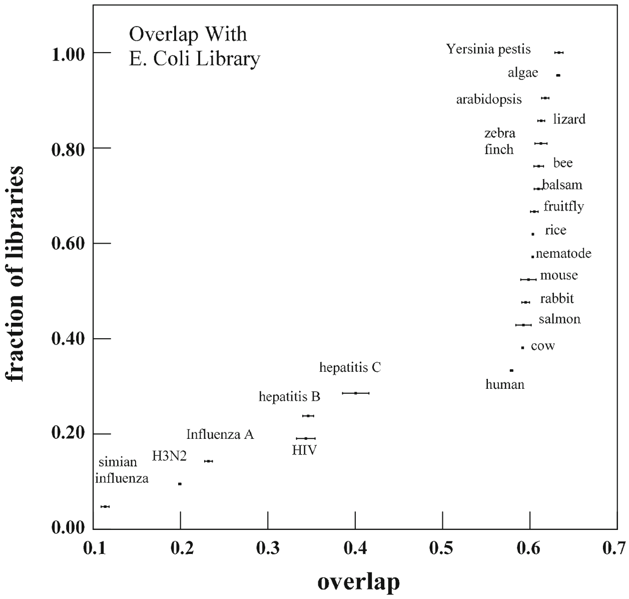


Fig. 7 Phase plot overlap with *E. Coli* libraries. The data are arranged *left*-to-*right* in order of increasing overlap. The *horizontal error bars* mark the average plus/minus one standard deviation

Phase plot differences can be quantified a second way via the correlations between $S_{\tilde G}$ and $S_{\tilde H}$. Where the correlations are substantial, precisely *how* the information accumulates in a protein significantly impacts the tuning modulations. Strong correlations within a phase plot signal that a large number of the library molecules are custom-designed. A *particular* value of $S_{\tilde G}$, say, 1.35 bits, demands a *specific* $S_{\tilde H}$, say, 2.37 bits, with no substitutions tolerated by the manufacturer. From the protein researcher's point of view, if he or she were to learn $S_{\tilde G}$ for a molecule in the library, the value of $S_{\tilde H}$ would arrive as a bonus.

The case of feeble correlations is different. Here the cumulative information minimally impacts the tuning modulations. A value of $S_{\tilde G}$ allows *several* possibilities for $S_{\tilde H}$, and vice versa. If $S_{\tilde G} = 1.35$ bits is paired with $S_{\tilde H} = 1.65,\ 3.05,\ 3.85$, and 4.15 bits, one size of $S_{\tilde G}$ accommodates several sizes of $S_{\tilde H}$. The protein manufacturer is not so choosy and exacting. And from a researcher's point of view, if he or she were to learn $S_{\tilde G}$ for a library entry, little information would be furnished about $S_{\tilde H}$.

As before, let $\Phi_A \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}$ quantify the fraction of points for library A in a small window of area $\Delta S_{\tilde G}\Delta S_{\tilde H}$. Along similar lines, let $\Phi_{A,\ S_{\tilde G}} \cdot \Delta S_{\tilde G}$ represent the fraction of points in the horizontal interval bounded by $S_{\tilde G}$ and $S_{\tilde G} + \Delta S_{\tilde G}$; let $\Phi_{A,S_{\tilde H}} \cdot \Delta S_{\tilde H}$ represent the fraction in $S_{\tilde H}$ and $S_{\tilde H} + \Delta S_{\tilde H}$. The pair-correlations *between* $S_{\tilde G}$, and $S_{\tilde H}$ can then quantified in *bits* via the mutual information [15]:

$$MI_A = \sum_{S_{\tilde G},S_{\tilde H}} \Phi_A \cdot \Delta S_{\tilde G}\Delta S_{\tilde H} \times \log_2\left[\frac{\Phi_A \cdot \Delta S_{\tilde G}\Delta S_{\tilde H}}{\left(\Phi_{A,S_{\tilde G}} \cdot \Delta S_{\tilde G}\right) \times \left(\Phi_{A,S_{\tilde H}} \cdot \Delta S_{\tilde H}\right)}\right] \tag{28}$$

Figure 8 illustrates results for the multiple protein sources. To discern the trends, the data have (again) been presented as a left-to-right ascendant distribution. As in the previous figure, error bars mark the average plus/minus one standard deviation. The errors are estimated by computing the mutual information for randomly-chosen samples of a population.

The trends are as striking as in Fig. 7. The mutual information is substantially greater for viral protein libraries, compared with living organisms. Regarding the latter, the collections engineered by nematodes (worms) demonstrate a slight correlation edge over fruit-flies and humans. Last place is occupied by the algae species. How

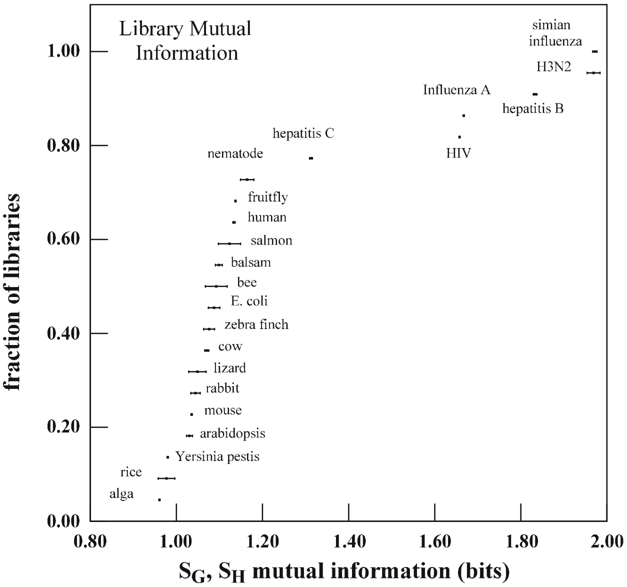


Fig. 8 Phase plot mutual information. The data are arranged *left-to-right* in order of increasing mutual information. The *horizontal error bars* mark the average plus/minus one standard deviation

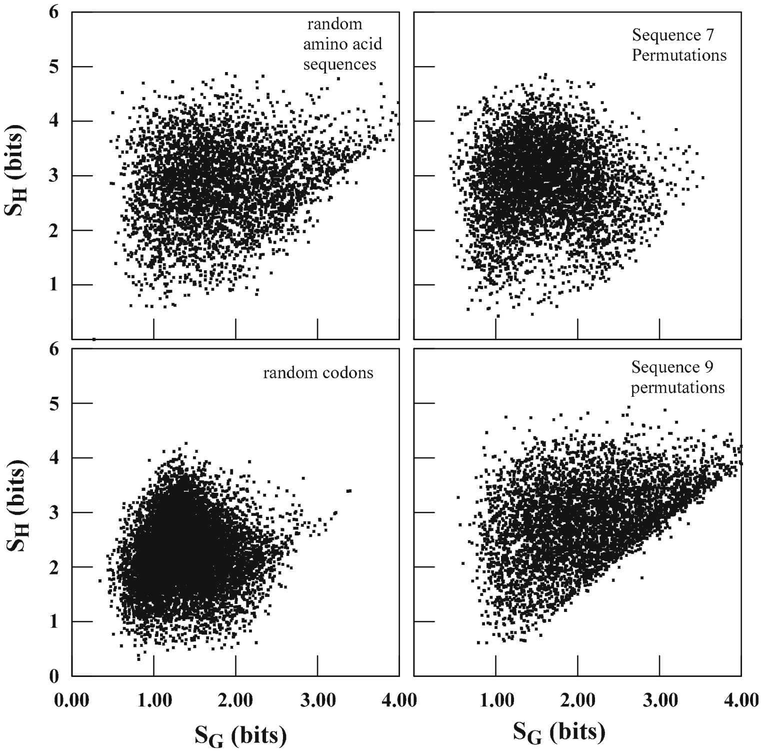
are viruses distinctive from living organisms? The answers include that viruses are superior in the custom-tuning of protein information. Figure 8 illustrates that living systems lag considerably behind.

The phase plots shown thus far derive from naturally occurring libraries. What if the libraries are not so natural? Figure 9 illustrates four examples. The phase plot in the upper left was constructed by formulating random strings of amino acids. While the string lengths were programmed to match those of *E. coli* proteins, there was no bias in the amino acid order or composition: V was placed during computations as often as nineteen other units. For a library of 5,000 molecules, the spectral complexity phase plot proves diffuse and exhibits none of the arrowhead and cleft motifs of natural sources.

The lower left panel applies to a random library as well. Here the precursors to proteins, namely polynucleotides, were assembled (virtually) with an absence of order and composition bias—the UCG codon (allied with serine) manifested as often as sixty-three other triplets. As is well known, the genetic code contains biases of its own, e.g. six codons pair with serine while only one codon is allied with methionine [16]. These and other inequalities led to phase plots that are less diffuse than for random amino acid strings. Yet not surprisingly, random codons do not capture the complexity morphology of natural protein libraries.

Along related lines, every amino acid sequence presents multiple permutations or sequence isomers. The right-hand panels of Fig. 9 show the phase plots obtained via permutations of (arbitrarily-assembled) Seqs. 7 and 9 of the

Fig. 9 Complexity phase plots for unnatural protein libraries. The *panels show* the results of applying four assembly algorithms described in text



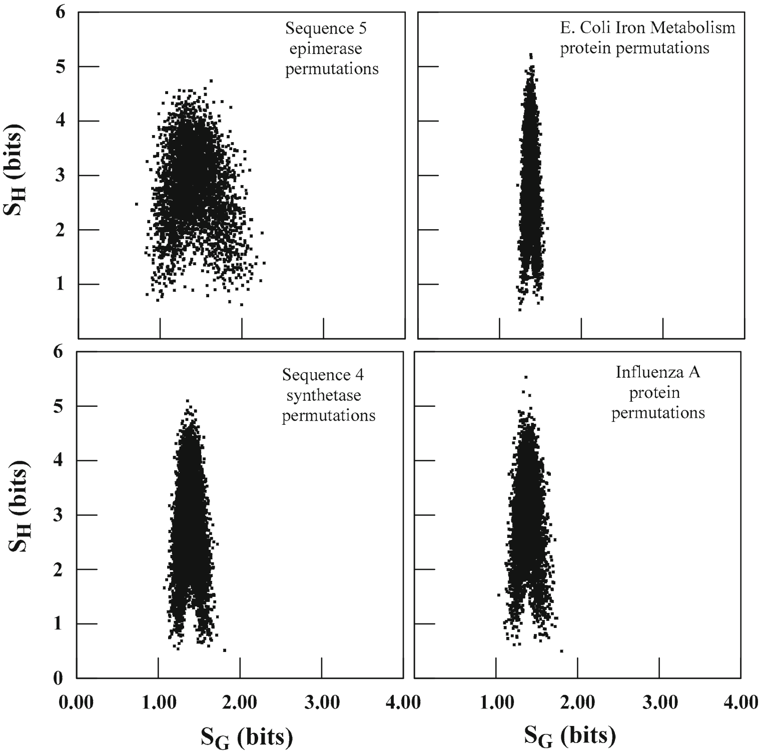


Fig. 10 Complexity phase plots derived from the primary structure isomers of natural proteins

Sect. 1. These panels further address the question: what is an unnatural protein? Apparently, such a molecule is incapable of generating the $S_{\tilde{G}}$, $S_{\tilde{H}}$ plots for natural libraries via re-orderings of the primary structure.

Figure 9 illustrates four (of infinite possible) ways *not* to simulate natural phase plots. What method succeeds and indeed sheds light?

Natural proteins are distinctive in that they contain information about one another. It is shown in the "Appendix", for instance, how the adenylosuccinate synthetase of Seq. 4 poses 10^{547} distinguishable rearrangements; a *very* large number of these, 10^{394}, spell out the epimerase of Seq. 5. Seq. 4 offers information about Seq. 5 in truly many respects.

The left-side panels of Fig. 10 show the state points placed by 5,000 random permutations of Seqs. 4 and 5. The upper right panel shows the phase plot obtained via permutations of the most complex (highest $S_{\tilde{H}}$ value) of *E. coli* proteins. The molecule is an iron metabolism protein which attains the highest-altitude in Fig. 4. The lower right panel contains the phase plot obtained via sequence permutations of a protein encoded by Influenza A. The patterns are striking in that *all* the panels evince the complexity motifs of natural libraries. A lesson is apparent. For any permutation an amino acid sequence, there is a relocation of the $S_{\tilde{G}}$, $S_{\tilde{H}}$ state point. For a naturally occurring molecule, the re-locations are restricted to the locus of complexity elected by nature. Permutation isomers thereby offer yet another way to distinguish natural from arbitrarily-constructed sequences. The latter are lacking in design rationales; in turn, they lack information which can reproduce nature's complexity. Interestingly, permutation isomers of virus-encoded proteins lead to $S_{\tilde{G}}$, $S_{\tilde{H}}$ plots appropriate to living organisms. The isomers of viral proteins *do not* reproduce the phase plots for either the source virus or other abiotics.

4 Discussion

Tanford and Reynolds [30] have referred to proteins as nature's robots. The description is fitting because the molecules control so much of nature's chemistry. Robots operate only according to their internal programming. Thus the letter sequences studied for this paper form the software for catalytic action and molecular signaling. As with all software, there are critical questions. Which programs are functional versus nonsensical? Which programs are less than benign for the host device? These questions are straightforward for everyday programs and computers. They are vexing for proteins, however, because the programming codes are so cryptic.

The results in the previous section are significant then because they establish additional assembly rationales of natural—and thus biologically operational—proteins. The rationales do not arrive via individual letter groupings per se. This is because the Fourier methodology weighs molecules by the totality of complexity; this property depends on *all* the amino acids that compose a compound. For a given system, information is cumulative and dispersed. Yet the systems designed and evolved by nature hew to a particular complexity in the mode structure. Moreover, the adherence proves universal for living sources and case-specific for viral, as readily fingerprinted via $S_{\tilde{G}}$, $S_{\tilde{H}}$ plots. It turns out to be easy to discriminate natural from arbitrarily-assembled compounds *without* engaging BLAST routines. Natural compounds offer $S_{\tilde{G}}$, $S_{\tilde{H}}$ in the arrowhead regions of Figs. 4, 5 and 6. The state points of arbitrarily-assembled molecules fall out of bounds.

The results also show natural proteins to carry information about one another. For a naturally-occurring molecule, permutations of the primary structure generate the $S_{\tilde{G}}$, $S_{\tilde{H}}$ distributions embraced by nature (cf. Fig. 10). Arbitrary sequences lack assembly rationales and provide only sparing, coincidental information about natural molecules. As would be expected (cf. Fig. 9), arbitrary sequences are unable to re-produce the phase plots for natural systems.

Living systems express $S_{\tilde{G}}$, $S_{\tilde{H}}$-distributions markedly different from viruses. This does not surprise. Both system types encode multiple proteins and one is as natural as the other. However, an organism encodes catalysts and signal carriers that support metabolism and capabilities such as immune responses. At life's threshold, a virus is neither metabolic nor defense-prone in the manner of an organism [17]. Thus the protein collections of viruses demonstrate glaring vacancies in the $S_{\tilde{G}}$, $S_{\tilde{H}}$-distributions along with amorphous boundaries. What proved interesting was the complexity divergence among viruses. Whereas the $S_{\tilde{G}}$, $S_{\tilde{H}}$ phase plots are similar for living systems (Figs. 4, 5), the plots for viral libraries prove quite the opposite. Note that the phase plots shown in this paper are evocative of blot patterns used in protein library analysis [20]. The plots for viruses (Fig. 6), subject to overlap and correlation analysis (Figs. 7, 8) prove signature-like, and enable ready assignment of the source.

The results are significant also because they identify new analytical tools for proteins. As is well-known, nature's catalysts are the frequent targets in drug development—many drugs are enzyme inhibitors [28]. Information spectra as in Fig. 2 and phase plots (Figs. 4, 5, 6) identify supplemental tools for the medicinal researcher in selecting targets. Inhibitors operate by modifying their target information; this disrupts the catalytic control of one or more reactions. Proteins that pose the maximum divergence information-wise from fellow library members should offer the minimum side-effects within the organism upon inhibition. The well-engineered inhibitor will impact the minimum number of catalysts.

Proteins encoded by viruses pose a sharp contrast. Here the catalysts that demonstrate the maximum information overlap should offer the most attractive targets. The drug synthesized at a pharmaceutical lab will impact the maximum number of catalysts and signal carriers. Both spectra and phase plots furnish a means of quantifying and contrasting the information divergence of target proteins.

4.1 Summary and Closing

A method for discriminating proteins based on molecular-level information has been presented. The method directs $Z_{CI}^{(i)}$-tables, Fourier analysis, and complexity phase plots to high-throughput advantage. Notably the complexity boundaries are universal for living systems and case-specific for viral libraries. Just as important, naturally-occurring molecules re-generate the phase plots of living systems via re-arrangements of the primary structure. As is well appreciated, natural protein libraries have a *real* capacity for replication: for living systems via mitosis and for viruses via host cell invasion. This capacity appears to coincide with the ability to generate complexity phase plots via sequence isomers. At present, the authors are delving further into the replication structure and symmetry of natural proteins. The results will be reported in a forthcoming paper.

Acknowledgments The authors appreciate discussions with Professors Miguel Ballicora and Ken Olsen about individual proteins, natural libraries, and NCBI archives. The authors wish to thank an anonymous reviewer for comments and suggestions for improvement.

Appendix

A natural protein contains information about another. Given two of nature's molecules, the primary structure of the smaller is generally contained in permutation isomers of the larger. We demonstrate how such ideas apply to Seqs. 4 and 5 of the Sect. 1. The approach is readily extended to other pairs of molecules.

Seqs. 4 and 5 are constructed from 457 and 120 amino acid units, respectively. The total number of permutations of (the larger) Seq. 4 is approximated as follows:

$$457! \approx \exp[457 \cdot \log_e 457 - 457] \approx \exp[2342] \approx 10^{1017}$$

Not all of the 10^{1017} are unique since **A** appears in 32 places, **V** in 40, **L** in 39, and so forth. By accounting for the placements of each amino acid, the number of distinguishable re-arrangements of Seq. 4 becomes:

$$\frac{457!}{32!\,40!\,39!\,25!\,18!\,18!\,7!\,9!\,52!\,19!\,26!\,5!\,12!\,17!\,20!\,25!\,27!\,30!\,26!\,11!} \approx \frac{10^{1017}}{7.96 \times 10^{469}} \approx 1.26 \times 10^{547}$$

Now let the amino acids of Seq. 4 be re-ordered so that the first (counting from left) 120 match Seq. 5. The result still allows a large number of permutations for “leftover” slots 121–457, namely:

$$\begin{aligned} &\frac{(457-120)!}{(32-10)!(40-8)!(39-10)!(25-14)!(18-2)!(18-2)!(7-2)!(9-2)!(52-1)!(19-5)!(26-8)!(5-0)!} \\ &\times \frac{1}{(12-4)!(17-9)!(20-5)!(25-8)!(27-9)!(30-6)!(26-10)!(11-5)!} \\ &\approx \frac{10^{705}}{1.84 \times 10^{313}} \approx 5.43 \times 10^{391} \end{aligned}$$

Seq. 4 can be re-arranged so that Seq. 5 commences at any one of $457-120+1$ slots. Thus the number of times Seq. 5 is contained in permutations of Seq. 4 is:

$$5.43 \times 10^{391} \times (457-120+1) \approx 1.84 \times 10^{394}$$

One sees that truly many permutation isomers of the larger molecule spell out the smaller one perfectly. A naturally-occurring protein carries structure information about another.

References

1. Agüero-Chapin G, González-Díaz H, de la Riva G, Rodríguez E, Sánchez-Rodríguez A, Podda G, Vázquez-Padrón RI (2008) J Chem Inf Model 48:434–448. doi:10.1021/ci7003225
2. Bajorath J (2000) Mol Divers 5:305–313. doi:10.1023/A:1021321621406
3. Bracewell R (1965) The Fourier transform and its applications. McGraw-Hill, New York
4. Chang G, Wang T (2011) Prot J 30:167–172. doi:10.1007/s10930-011-9318-0
5. Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN (2008) Chem Res Toxicol 21:619–632. doi:10.1021/tx700296t
6. Glisic S, Veljkovic N, Cupic SJ, Vasiljevic N, Prljic J, Gemovic G, Perovic V, Veljkovic V (2012) Prot J 31:129–136. doi:10.1007/s10930-011-9381-6
7. González-Díaz H, Dea-Ayuela MA, Pérez-Montoto LG, Prado-Prado FJ, Agüero-Chapín G, Bolas-Fernández F (2009) Mol Divers. doi:10.1007/s11030-009-9178-0
8. González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007) J Proteome Res 6:904–908. doi:10.1021/pr060493s
9. Graham DJ, Greminger JL (2009) Mol Divers. doi:10.1007/s11030-009-9211-3
10. Graham DJ, Greminger JL (2011) Mol Divers. doi:10.1007/s11030-011-9307-4
11. Graham DJ, Malarkey C, Schulmerich MV (2004) J Chem Inf Comput Sci 44:1601–1611. doi:10.1021/ci0400213
12. Graham DJ, Pilarski B (2000) J Phys Chem B 104:329–341. doi:10.1021/jp992107i
13. Iancu CV, Borza T, Fromm HJ, Honzatko RB (2002) J Biol Chem 277:40536. doi:10.1074/jbc.M204952200
14. Jiang L, Li M, Wen Z, Wang K, Diao Y (2006) Prot J 25:241–249. doi:10.1007/s10930-006-9007-6
15. Kullback S (1997) Information theory and statistics, Chap 1. Dover, New York 1
16. Lehninger AL (1970) Biochemistry, Chaps 30 and 10. Worth Publishers, New York
17. Levy JA, Fraenkel-Conrat H, Owens RA (1994) Virology, Chap 1. Englewood Cliffs, NJ
18. Liu H, Yang J, Wang M, Xue L, Chou K-C (2005) Prot J 24:385–389. doi:10.1007/s10930-005-7592-4
19. MacDonald DK (2006) Noise and fluctuations, Chap 2. Dover, New York
20. Müller HJ, Röder T (2006) Microarrays. Academic Press, Waltham, MA, Chapter 3
21. Muraki M, Harata K, Sugita N, Sato K (1996) Biochem 35:13562. doi:10.1021/bi9613180
22. Nakanishi K, Fukai S, Ikeuchi Y, Soma A, Sekine Y, Suzuki T, Nureki O (2005) Proc Nat Acad Sci USA 102:7487–7492. doi:10.1073/pnas.0501003102
23. Ploom T, Haussmann C, Hof P, Steinbacher S, Bacher A, Richardson J, Huber R (1999) Structure Fold Des 7:509–516. doi:10.1016/s0969-2126(99)80067-7

24. Prado-Prado F, García-Mera X, Abeijón P, Alonso N, Caamaño O, Yáñez M, Gárate T, Mezo M, González-Warleta M, Muiño L, Ubeira FM, González-Díaz H (2011) Eur J Med Chem 46: 1074–1094
25. Raines RT (1998) Chem Rev 98:1045–1066. doi:10.1021/cr960427h
26. Randić M, Zupan J, Balaban AT, Dražen V-T, Plavšić D (2011) Chem Revs 111:790–862. doi:10.1021/cr800198j
27. Riera-Fernández P, Munteanu CR, Escobar M, Prado-Prado F, Martín-Romalde R, Pereira D, Villaba K, Duardo-Sánchez A, González-Díaz H (2012) J Theor Biol 293:174–188. doi:10.1016/j.jtbi.2011.10.016
28. Robertson JG (2005) Biochem 44. doi:10.1021/bi050247e
29. Stahura FL, Godden JW, Xue L, Bajorath J (2000) J Chem Inf Comput Sci 40:1245–1252. doi:10.1021/ci0003303
30. Tanford C, Reynolds J (2001) Nature’s Robots: a history of proteins. Oxford University Press, Oxford
31. Vogt I, Ahmed HEA, Auer J, Bajorath J (2008) Mol Divers 12:25–40. doi:10.1007/s11030-008-9071-2