Mol Divers (2011) 15:769–779
DOI 10.1007/s11030-011-9307-4

FULL-LENGTH PAPER

On the information expressed in enzyme structure: more lessons from ribonuclease A

Daniel J. Graham · Jessica L. Greminger

Received: 8 September 2010 / Accepted: 5 February 2011 / Published online: 24 February 2011

Abstract Brownian computations were directed at Ribonuclease A (RNase A) and variants in folded states so as to quantify information of the statistical type at the atom/covalent bond level. This advanced the research reported in this journal last year on the information properties of enzyme primary structure. Brownian computation data are illustrated for a sixteen-member library. The results identify signature traits that distinguish the folded wild type (WT) molecule from variants. The distinctions are explainable in terms of correlated information and dispersion energy. The Brownian tools used for this study can be directed at other protein families (e.g., kinases, isomerases, etc.) in rapid screening, QSAR, and design applications.

Keywords Enzyme structure · Enzyme function · Enzyme information · Brownian computation · Molecular information

Abbreviations

PDB	Protein data bank
RNase A	Ribonuclease A
RNA	Ribonucleic acid
QSAR	Quantitative structure activity relation
CI	Correlated information
MI	Mutual information
WT	Wild type
ABA	Atom-bond-atom
1D	One dimensional
3D	Three dimensional

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

Introduction

The 124-residue amino acid sequence for bovine wild type Ribonuclease A (WT RNase A) can be represented most succinctly as [1]:

KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNL
TKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQT
NCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHII
VACEGNPYVPVHFDASV

Like all proteins, the molecule offers a large number of substitution possibilities. Each member of the sequence permits 19 alternates given the standard 20 amino acids. Thus the number of possible single-site variants Ω_1 is:

$$\Omega_1 = 19 \cdot 124 = 2356 \tag{1}$$

Regarding double-site variations, the N-terminal moiety K (lysine) permits 19 alternates; the 123 sites to its right (ETAAA…) each allow 19. The second-from-left site E (glutamic acid) permits 19 substitutions and the same is true for the next 122, and so on. Such considerations point to Ω_2, the number of possible double-site variants:

$$\begin{aligned}\Omega_2 &= (19 \cdot 19 \cdot 123) + (19 \cdot 19 \cdot 122) + (19 \cdot 19 \cdot 121) \\ &\quad + \cdots + (19 \cdot 19 \cdot 3) + (19 \cdot 19 \cdot 2) + (19 \cdot 19 \cdot 1) \\ &= 19 \cdot 19 \cdot (123 + 122 + 121 + \cdots + 3 + 2 + 1) \\ &\approx 2.75 \times 10^6 \end{aligned} \tag{2}$$

As for triple-site variations, the number possible must be a multiple of 19^3. The ideas underpinning Eqs. 1 and 2 can be combined to yield Ω_3:

$$\begin{aligned}\Omega_3 &= 19 \cdot 19 \cdot 19 \cdot [(122 + 121 + 120 + \cdots + 2 + 1) \\ &\quad + (121 + 120 + \cdots + 2 + 1)\end{aligned}$$

$$+ (120 + 119 + \cdots + 2 + 1) + \cdots]$$
$$\approx 2.13 \times 10^9 \quad (3)$$

Clearly, the number of variations in the protein formula increases exponentially with amino acid substitutions. Whether WT or variant, the sequence dictates the three-dimensional (3D) folded structure along with the chemical function.

Now regarding the folded structure, Fig. 1 illustrates the backbone of WT RNase A and one example each of a mono-, di-, and tri-substituted variant: *H12A, C40A/C95A*, and *K7A/R10A/K66A* [2–5]. Open circles mark the substitution regions while Eqs. 1–3 quantify all the variant possibilities. If the folded structures of all possible mono-, di-, and tri-substituted variants were available, one would have to reckon with over two billion panels of the Fig. 1 genre. However rich, protein data banks (PDBs) archive information for only a small fraction of possibilities, now and in the foreseeable future.

The similarities of the Fig. 1 samples are striking and yet typical. The structure differences are subtle and fleshed out only by point-by-point comparisons of the amino acid side-chain orientations and packing distributions. Clearly, the variants demonstrate substantive conservation of the WT features, in particular the carbon and amide backbone. It is reasonable to expect that thousands or more variants, if prepared and probed by X-ray diffraction, would demonstrate closely aligned configurations.

What is special about the WT *besides* that it is the one selected by nature? There are well-established answers, of course, both chemical and physical. The WT and variants share the capacity for folding as in Fig. 1 and catalyzing RNA hydrolysis. The WT is distinctive, however, for its stability, hydrolytic site specificity, and level of activity [6]. Variants, especially those with amino acid substitutions at binding and active sites, demonstrate lesser ability on one to several accounts. Using bench methods—activity measurements, mass spectrometry, sequencing procedures, etc.—a chemist would be well empowered to distinguish a WT sample from a collection of variants. Yet because of the backbone conservation and other 3D similarities, predictions based on casual structure inspections alone are far less clear. One would be hard-pressed, for instance, to pick out the WT molecule juxtaposed with variants. For example, only with

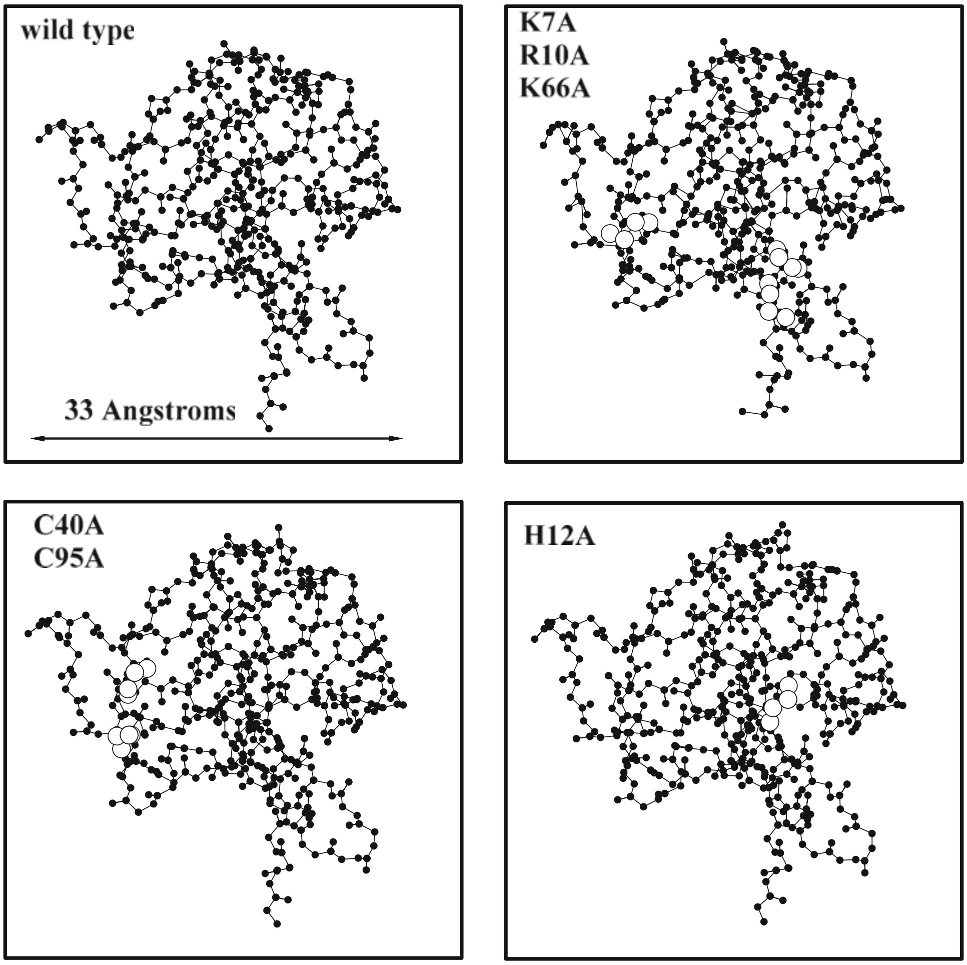

Fig. 1 Backbone structures of WT RNase A (PBD 1FS3), *K7A/R10A/K66A* (PDB 3RSK), *C40A/C95A* (PDB 1A5P) and *H12A* (PDB 1C9V). *Open circles* mark the main-chain atoms of substituted residues. The backbone structure of the WT is conserved to significant degree upon substitution

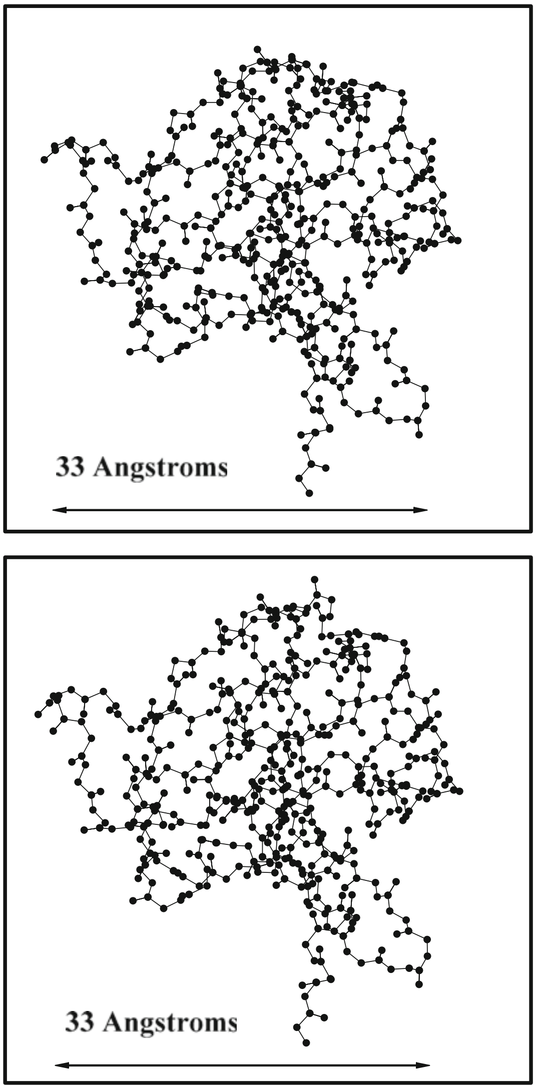

Fig. 2 Backbone structures of WT and variant RNase A. Which is which? The answer appears in the text

PDB-access and alignment software could a chemist ascertain which of the Fig. 2 panels corresponds to the WT and which to a variant. As it turns out, the lower panel illustrates one of the 2356 possible single-site variants, namely T45G; this molecule demonstrates significantly less hydrolytic site discrimination compared with the companion molecule [7]. Given the alignment similarities, what characteristics make the upper panel compound so special?

Questions of similarity and singularity are frequently addressed in chemistry [8]. To cite examples, Bajorath et al. [9–11] have investigated the characteristics that distinguish naturally occurring from synthetic compounds. The research has been motivated to great extent by therapeutic applications; it is important to know how to design molecules so that they mimic more closely the ones prepared by nature. Likewise, Sadowski, Kubinyl and Ajay et al. [12,13] have explored the signature traits of drugs versus non-drugs. A better understanding in this area offers strategies for maximizing drug potency and minimizing side-effects. Different proteins have dedicated catalytic functions: hydrolysis, oxidation/reduction, etc. González-Díaz et al. [14] have established the Markov and entropic characteristics of the substantive family of kinases. More pertinent to this study, González-Díaz et al. have quantified the molecular diversity, alignment, and mass fingerprint characteristics of the ribonuclease family [15,16]. Along related lines, McGaughey et al. [17] have investigated the means for ranking—quantifying the specialness—of kinase inhibitors. Since a substantial number of kinases are involved in diseases, understanding their chemical function reveals both drug targets and structure leads for control ligands. The specialness of molecules impacts two ways—beneficial and harmful. In the latter case, researchers have quantified the electronic aspects of hydrocarbons and nitrosamines that underpin carcinogenicity and toxicity [18–20].

In the cited efforts and more, the overriding themes are similarity combined with specialness. Why is nature attuned to certain compounds, when the variant possibilities are large in number with subtle structural differences?

In a previous study, the authors inquired about the specialness of the primary structure of active proteins, using RNase A as a test case [21]. The signature traits were less than clear-cut given that the building blocks (amino acid units) allow many sequence variations, all with the same mass, charge, and functional group diversity. It was demonstrated that the native sequence is distinguished, although not uniquely, by an economy of information. The choice of amino acid residues by nature was skewed toward the ones expressing low information of the statistical type in the atom/covalent bond networks. The native sequence further displayed a linear scaling of the cumulative information. This meant that the amino acid order of internal fragments of RNase A such as …*SRNLTKDRCK*… is strongly correlated with the order of fragments throughout the system. Nature's choice of the cysteine locations and disulfide bonds only reinforces these correlations. This linear scaling and the bias toward low information residues were shown to be salient traits of bioactive proteins in general, not just RNase A.

This article delves more deeply into information of the statistical type for active proteins, and looks again to RNase A for lessons. This was feasible given PDB archives and the adaptability of Brownian computations to folded systems. The findings are noteworthy because they show that the information effects of amino acid substitutions to be *not* localized but rather dispersed across a folded protein. Further, the WT molecule is distinctive both for its enhanced mutual information (MI) and low dispersion of covalent bond enthalpy relative to variants. The explanations are straightforward and identify new strategies for tuning the chemical diversity and activity of an enzyme. The Brownian tools used for this study can be directed at other protein families (e.g., kinases, isomerases, etc.) in rapid screening, QSAR, and design applications. These complement the entropic descriptor approaches established for enzymes which include RNase A by previous researchers [14–16].

Enzymes and Brownian computations

Enzymes were the first molecules examined as Brownian computers. Bennett [22] discussed how polymerases exercise nearest-neighbor random walks that are forced along one direction of DNA. The actions include the trapping of electronic information at individual base sites and the synthesis of complementary strands. Bennett used the term *Brownian* because the information processing transpires erratically and haphazardly via thermal collisions. Importantly, Brownian computation is not confined to polymerases and DNA. Information is that which can alter a probability distribution [23]. Every compound, enzyme and otherwise, carries information in an atom/covalent bond network and communicates by electronic contacts made in solution.

Molecules are electric charge assemblies that admit coding by formula diagrams. The constituents of the diagrams are familiar atom-bond-atom (ABA) units: *C–C*, *C=C*, *C–O*, *C=O*, …. These form an alphabet or digital code for communicating chemical functions, QSARs, and reactions. Using this alphabet to encode thermal collision sequences has been the focus of several studies from this lab [21,23–26]. The sequences allowed by a molecule have the form of electronic record tapes, e.g.,

…(C–H)(C–H)(C–C)(C–C)(C–H)(C–C)(C–H)(C–C)
(C–H)(C–H)(C–C)(C=C)(C–H)(C–C)(C–H)(C–H)
(C–C)(C=C)(C–C)(C=O)(C–C)(C=C)(C–H)(C=C)
(C–H)(C=C)(C–C)(C–H)(C–C)(C–H)(C–C)(C–H)
(C–C)(C–H)(C–H)(C–H)(C–H)(C–H)(C–C)(C–C)
(C–H)(C=C)(C–C)(C–H)(C–H)(C–C)(C=C)(C–C)
(C=C)(C–C)(C–C)(C–C)(C–C)(C–H)(C–C)(C–C)…

The electrical nature of a molecule is captured by its ABA diagram. The electrical messages, in turn, are embodied by the possible ABA sequences accessed by collisions. When it comes to proteins, all systems pose *facts-and-data* information at the various structure levels; this is the substance of PDBs. Quantifying information of the *statistical* type for folded systems was the principal objective of this project. Our foundation study distinguished important differences between the primary structure of WT RNase A and sequence isomers [21]; this study considers the more complex folded states of the protein and variants.

The authors focused on the sixteen molecules listed in Table 1. All constitute RNase A with complete structures solved by X-ray diffraction at comparable resolution. Figures 1 and 2 portrayed four of the sixteen. If the remaining twelve were included, the viewer, lacking a scorecard, would find it difficult to match the graphics correctly with Table 1 entries. All sixteen proteins exhibit substantial backbone alignment with disparities due largely to the amino acid side-chain orientations and packing distributions. Fifteen variants are far fewer than allowed by Eqs. 1–3. Even so, they cover a wide range of chemical effects. *C[40, 95]A* derives from *S–S* excision and twin *A*-substitutions which have modest effect on the enzyme activity [4]. By contrast, *H12A* and *H119A* demonstrate $> 10^4$ activity diminution because of changes at the active site [3]. *T45G* (cf. Fig. 2) demonstrates a marked decrease in substrate site-specificity and enhanced catalytic activity as a result [7]. Substitutions at *Phe46* and *Phe120* confer significant activity changes, whereas the sensitivity at *Pro93*, *Lys7*, *Arg10*, and *Lys66* is less acute [2,4,5,27,28]. *Asp121* is classified as an active-site residue. *D121A* demonstrates sparse activity reduction, however, compared with the WT [29]. The point is that a spectrum of bio-catalytic effects is represented in Table 1. The reader is directed to the cited literature for detailed accounts.

Table 1 Library of RNase A molecules investigated by Brownian computations

RNase A (PDB ID)	References
Wild type (1FS3)	[2]
F120A (1EIC)	[2]
F120G (1EID)	[2]
F120W (1EIE)	[2]
H12A (1C9V)	[3]
H119A (1C9X)	[3]
C40A, C95A (1A5P)	[4]
P93A (1A5Q)	[4]
K7A/R10A/K66A (3RSK)	[5]
T45G (1C8W)	[7]
P93G (3RSP)	[28]
D121N (3RSD)	[29]
D121A (4RSD)	[29]
F46L (1IZP)	[27]
F46V (1IZQ)	[27]
F46A (1IZR)	[27]

The PDB archives enabled encoding of the RNase A structure diagrams in 3D virtual formats. Our small-molecule studies had used adjacency matrices for encoding [24–26]. For proteins, however, compressed formats are necessary using integer arrays of dimension $6A$: A equals the number of the atoms excluding hydrogen [21]. For this project, the row entries specified the sites in each structure diagram along with the atom and residue identities, plus all the covalent linkages. Computational programs written by the authors interfaced with the PDB files and logged the atomic coordinates

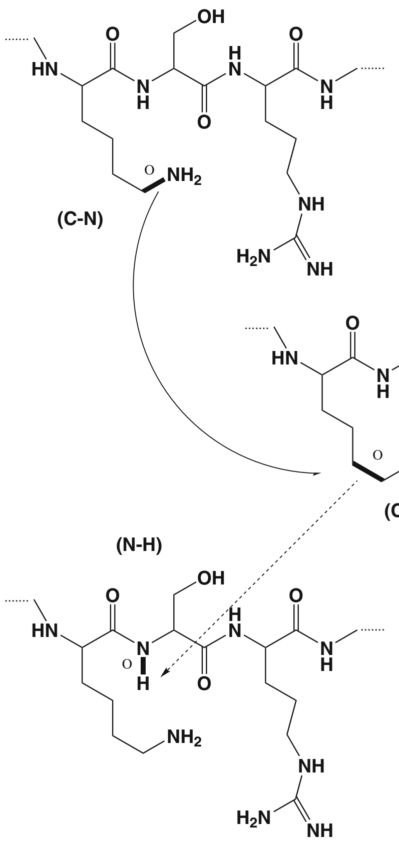

Fig. 3 Algorithm for compiling electronic message tapes. For a folded protein, a Brownian walk is a succession of nearest-neighbor ABA jumps by an atom such as helium (*open circle*); the jumps are both covalent link and through-space in nature. The *upper two diagrams* represent an example of a nearest-neighbor covalent-link jump: $(C-N)_{\mathrm{Lys31}} \rightarrow (C-C)_{\mathrm{Lys31}}$. The *middle* and *lower-most diagrams* show an example of a nearest-neighbor through-space jump: $(C-C)_{\mathrm{Lys31}} \rightarrow (N-H)_{\mathrm{Ser32}}$. The identity of the ABA unit is registered and recorded with each jump. The spatial coordinates are provided by X-ray structure data

in floating point arrays. The positions of the hydrogen atoms were inferred based on C, N, O, and S valence rules, pH, and excluded volume properties. The methods were largely equivalent to the ones used in our investigation of enzyme primary structure [21]. The key difference was that X-ray data governed the placement of the atoms. The enzymes were investigated for their intrinsic information properties, that is, independent of substrate, inhibitor, and solvent molecules.

Enzymes are diverse macromolecules. Even so, their structure diagrams can be encoded using a minimalist palette: *C–C*, *C=C*, *C–O*, *C=O*, *C–N*, *C=N*, *C–S*, *C–H*, *N–H*, *O–H*, and *S–S*. When compiling the Brownian record tapes, the random walk sequences of these units were established. To conserve computer memory, the ABA-units were abbreviated by single letters *A*, *B*, …, *K*. The algorithm used for assembling the tapes is summarized in Fig. 3.

The figure illustrates that if a diffusing atom, say, helium indicated by the open circle, were to collide with *(C–N)* of Lys31 of RNase A, a succeeding collision may involve *(C–C)* of the same residue and covalent link. The single step portrayed by the upper two diagrams is accordingly:

$$(C-N)_{\mathrm{Lys31}} \rightarrow (C-C)_{\mathrm{Lys31}}$$

Yet the translational accessibility of ABA units in the neighborhood of the walker must be taken into account, e.g., *(C=O)*, *(C–N)*, etc. of *Ser32*. These units are situated at coulombic, hydrogen bond, and van der Waals distances over a 3–5 Å range. Thus the middle and lower-most diagrams of Fig. 3 illustrate an example of a through-space jump:

$$(C-C)_{\mathrm{Lys31}} \rightarrow (N-H)_{\mathrm{Ser32}}$$

In a folded protein, a random walk is a succession of nearest-neighbor jumps, both covalent link and through-space: $\ldots(C-N)_{\mathrm{Lys31}} \rightarrow (C-C)_{\mathrm{Lys31}} \rightarrow (N-H)_{\mathrm{Ser32}} \rightarrow (N-C)_{\mathrm{Ser32}}\ldots$

Each step registers an ABA site, as with a collision in a thermal environment. The coordinates of the impact site are the ones located by X-ray diffraction. In Brownian computations, following a random walk step, a subroutine scans the structure file so as to identify the jump-accessible sites. A structure file based on diffraction data does not quantify electronic interactions per se. As a consequence, the Brownian algorithm treats all collision-eligible sites as equally so. If ten sites are accessible to the walker, the probability of each is taken as 0.100. One of the ten is then selected at random. The critical feature is that the possible steps and sequences are governed from beginning to end by the folded protein structure.

When written in abbreviated formats, a message tape for WT RNase A looks like:

…ADDDDEDEEEDDAHDDEDDAAAHADEEADADDD
ADEEIEEDEDDCHADEEDHHADAAEDIHDADEIEDA
DHAAAEHIEACEDDJDEDEADAHEDDDDAIDDEIADI
HAEDDAAEEDEDADDDAEDHAAIDADDIHEDDDAEI
EAAACHDEADIAEIAEDDEJHDEIDADAEEDDDEAA
DEDEEAEDAHDAAEHEDIEDHAEAADEAEDHAAEAJ
AEEEHAHEIDDADEDEEHAAHIDCDADEHEDHHDD
DIAEDDDEHDDDEDHDAADEEIEEDHAHEADAAED
DAEDAADEEDDEHADDEDEDEEEDEDEHIEAIAHAE
HDEAADDEAADHEDDDDDDDDDIIIAIEAADDEDD
EIDEDHAAADADDDDDEIDADEADEHEIAHDDEIDE
EEJEEDEAHEDDEACAEAIAAAEAAADEHEDAEIHD
DDEADDAIDHEADDADDADAEDHIEAADEDJDEADI
DAEIEADEADEAAHEEDEAHHDDHDDEHDCDIAEA
DEDECDEDEIDCAADEEJDEDEAEDEDEAHACEEAD
DEDEDDAEDADEEEADDIAEAEEDCAEDEEDIDAEE
IDA…

However nonsensical, the tape details the electronic sequences accessed during a random walk; in other words, the electronic messages available from the protein in a thermal environment. Records of seven million units were compiled for each system of Table 1. This ensured walker step/ABA ratios of approximately 3600:1 and strict convergence of the record tape energy and information properties. All the collision sites were thoroughly sampled in the computations. No new information was realized using larger walker step/ABA ratios.

The record tapes were parsed for their nth-order states represented by the code strings:

$$\begin{array}{ll} A, B, C, \ldots & n = 1 \\ AA, AB, AC, \ldots & n = 2 \\ AAA, AAB, AAC, \ldots & n = 3 \\ \vdots & \end{array}$$

The occurrence frequencies $f_i^{(n)}$ were subsequently quantified along with enthalpy values $E_i^{(n)}$. For example, the third-order state $AAC \leftrightarrow (C-C)\ (C-C)(C-O)$ was paired with

$$E^{(n=3)} = 348 + 348 + 355\,\mathrm{kJ/mol} \tag{4}$$

The energy terms were supplied by standard look-up tables [30]. The record tapes were assessed for the enthalpy moments $\langle E^{(n)} \rangle$ and $\sigma^2_{E(n)}$. The square root of the latter is the standard deviation and measure of enthalpy dispersion.

A molecule's 3D structure—the ordered configuration of atoms and chemical bonds—determines the chemical function. Thus, the project concentrated on the correlated or MI contained in the record tapes. If nature assembled proteins and other molecules without structure rules, their message tapes would express zero MI. Fortunately this is not the case.

At each nth order of analysis, the message correlations were quantified via the formulae:

$$\mathrm{MI}^{(n=2)} = \sum_{i,j} f_{ij} \log_2 \left(\frac{f_{ij}}{f_i \cdot f_j} \right) \tag{5}$$

$$\mathrm{MI}^{(n=3)} = \sum_{i,j,k} f_{ijk} \log_2 \left(\frac{f_{ijk}}{f_i \cdot f_j \cdot f_k} \right) \tag{6}$$

$$\vdots$$

up to the eighth order. Note that MI is enhanced by message tape fragments such as

…ABABCABABC…
…CBAACBAACB…
…DEAGAGDEAG…

This is because the repeating patterns offer information about the collision sites logged down-stream on the record tape. These patterns reflect the structure constraints and local symmetry within the molecule. By contrast, MI is diminished by fragments such as

…ACDGIJCFAH…
…CCKHEDGMAB…
…ALBHEDAGCD…

Because of their disorder, fragments of this sort offer little information about downstream messages. What units (collision sites of the molecule) follow H, B, and D in the above? The answers are less than obvious and reflect the asymmetrical spatial attributes of a protein.

The enthalpy dispersion σ_E is enhanced by message tape fragments (written using non-abbreviated symbols) such as

…(C–C)(C=C)(C–S)(O–H)(C=N)(C–S)(C=N)(S–S)…

The ABA enthalpies span a range of 250–614 kJ/mol. By contrast, σ_E is narrowed by fragments such as

…(C–C)(C–O)(C–C)(C–N)(C–S)(S–S)(S–H)(C–S)…

where the range is 259–355 kJ/mol. As the project found, the WT and variants diverge on *both* MI and σ_E accounts. In particular, nature prefers an enzyme with enhanced correlations of electronic messages, and lower σ_E relative to variants.

This investigation detailed how MI and σ_E were distributed spatially in RNase A, following the procedure of the previous research [21]. Briefly, a spherical barrier was programmed at each residue site indexed by integer λ. RNase A contains 124 residues, so λ runs from 1 to 124. A record tape was compiled for the atom and covalent bond configuration within each barrier. Naturally, several residues contributed

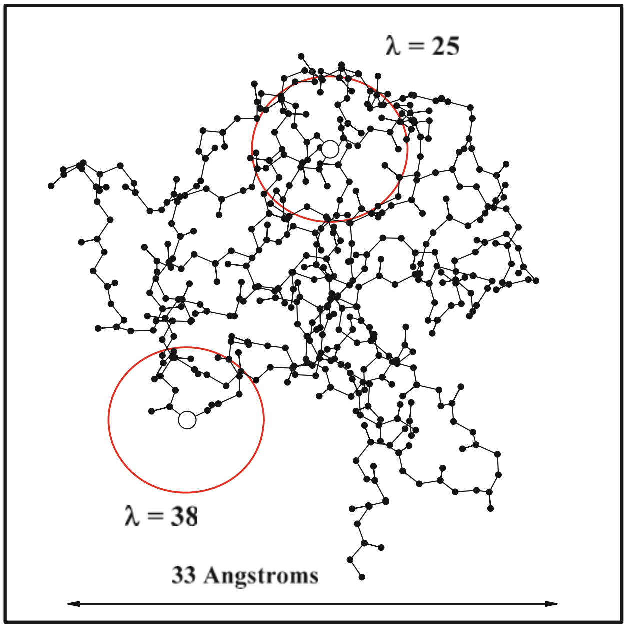

Fig. 4 Information distribution in a folded protein. A spherical barrier of radius 5 Å was programmed at each neighborhood indexed by integer λ. The *large circles* portray mark barriers at λ = 25 and 38 where the α-carbons of *Tyr25* and *Asp38* lie at the centers. The Brownian computations establish the electronic messages of the configurations within each barrier

to each configuration. Figure 4 illustrates two examples: the large circles mark virtual barriers at $\lambda = 25$ and 38 whereby the α-carbons of *Tyr25* and *Asp38* lie at the centers (small open circles). The configuration of ABAs within the barriers is dictated entirely by the primary, secondary, and tertiary structure of the protein.

124 barriers were programmed for each Table 1 system. Brownian record tapes, and MI, σ_E analyses were compiled at length. For conciseness, the data shown here are confined to computations based on equal-sized barrier-regions of radius 5 Å . The use of much larger radii blurred the information effects, while smaller ones inadvertently excluded ABA sites.

Results

What proved most significant were the *regio*-distributions of MI and σ_E. Figure 5 shows the results for WT RNase A: the peaks and valleys of MI and σ_E proved not concentrated in any one region—the active, binding, and recognition sites are neither more nor less endowed with message correlations. What was striking were the fluctuations: MI varies up to 50% across the protein while σ_E fluctuates 20–25%. The information and enthalpy fluctuations significantly exceeded the errors imposed by the random walk algorithm. These errors have been indicated in Fig. 5 by the vertical bars.

A protein's structure correlations and energy are altered by amino acid substitutions. Figure 6 shows the distribution *differences* between *K7A/R10A/K66A* and WT RNase A:

$$\Delta \mathrm{MI}_\lambda = \mathrm{MI}_\lambda^{(K7A/R10A/K66A)} - \mathrm{MI}_\lambda^{(\mathrm{WT})} \quad (7)$$

$$\Delta\sigma_{E(\lambda)} = \sigma_{E(\lambda)}^{(K7A/R10A/K66A)} - \sigma_{E(\lambda)}^{(\mathrm{WT})} \quad (8)$$

As before, error bars mark the averages ± one standard deviation. The differences in MI and σ_E occur in both positive and negative directions. At the same time, the error bars attest that not all the changes are significant. This means that the amino acid substitutions modify the protein information in a selective and dispersed way. In the case of Fig. 6, the most substantial changes are apparent near *Lys7* while comparable changes manifest at sites spatially well removed.

Statistical significance was assigned to the differences that exceeded three standard deviations. Applying this standard led to plots such as Fig. 7. Here the WT molecule has been illustrated again. Open circles mark the regions where statistically significant ($\geq$ 99% confidence level) changes in MI were effected by substitution. Figures 6 and 7 follow specifically from comparing WT and *K7A/R10A/K66A*. Information-wise, the substitutions impact the neighborhoods surrounding 27 residues, i.e., 24 more than the number of substitution sites. This shows the degree to which the information effects are dispersed, not localized, in the protein. The same type of plot can be constructed to illustrate the changes in σ_E.

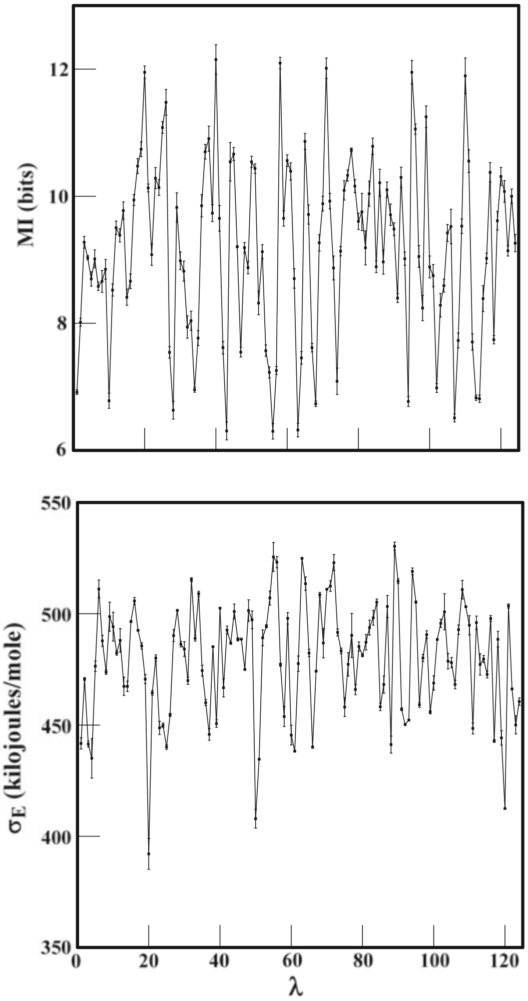

Fig. 5 Distribution of information and enthalpy dispersion. *Upper* and *lower panels* respectively show MI and σ_E for WT RNase A as a function of region index λ. The *error bars* mark the average ± one standard deviation

An enzyme's 3D structure can be viewed from innumerable perspectives simply by rotating the image. Fifteen variants were examined for the project. Each offers definitive point-by-point contrasts with the WT and with one another. Quantifying the information for a modest-size enzyme library yields an abundance of data.

To condense and simplify matters, the significant MI and σ_E differences were summed for each variant. The results then identified the location of a state point for each system. The state space can be represented in a plane defined by the differences with respect to the WT. The ordinate and abscissa of the plane correspond to $\sum_\lambda \Delta\mathrm{MI}_\lambda$ and $\sum_\lambda \Delta\sigma_{E(\lambda)}$, respectively. The result is Fig. 8 whereby the open square marks the WT point at 0.00 bits, 0.00 kJ/mol. The points for the variants appear in reference to the origin while the error bars mark the average ± one standard deviation. If the effects of amino acid substitutions were insig-

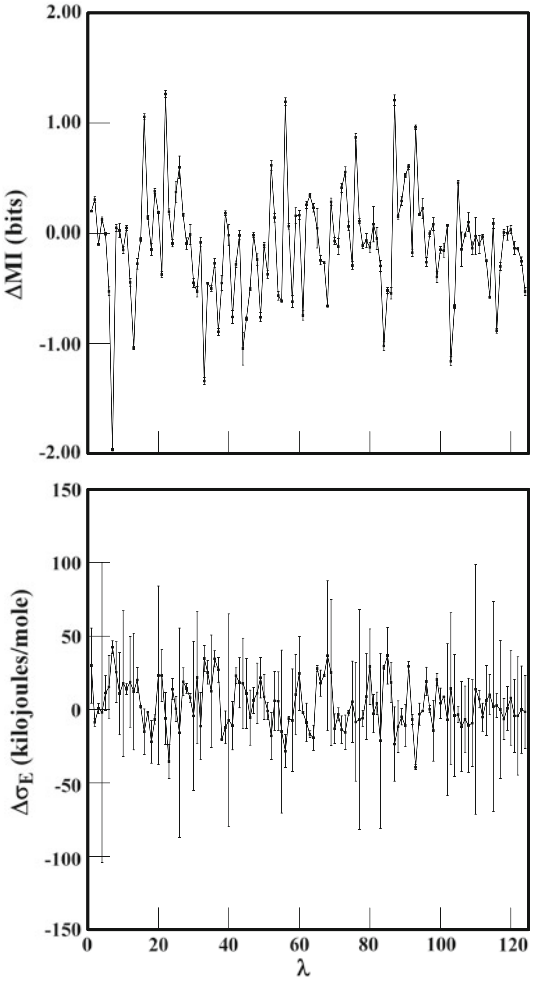

Fig. 6 Differences between native protein and variant. Plotted are $\Delta\mathrm{MI}_\lambda$ and $\Delta\sigma_{E(\lambda)}$ of Eqs. 5 and 6 as a function of region index λ. The variant corresponds to *K7A/R10A/K66A*. *Error bars* mark the averages ± one standard deviation

nificant, all the state points would be clustered at the origin. This is not the case.

The points for variants are found at *negative* $\sum_\lambda \Delta\mathrm{MI}_\lambda$. Evidently the electronic message correlations for the WT *exceed* those of variants. Concomitantly, the variants demonstrate *positive* $\sum_\lambda \Delta\sigma_{E(\lambda)}$ with one exception, *C[40,95]A*. This means that, by and large, the WT bond enthalpies manifest a narrower distribution compared with variants.

It is instructive to examine the likelihood of observing such results, assuming pure chance. Amino acid substitutions in a protein either enhance or diminish MI; likewise for σ_E. For a variant population of size N, the probability of observing N_+ number of positive effects (enhancements) is

$$\mathrm{prob}(N_+) = \frac{\frac{N!}{N_+!(N-N_+)\,!}}{2^{15}} \qquad (9)$$

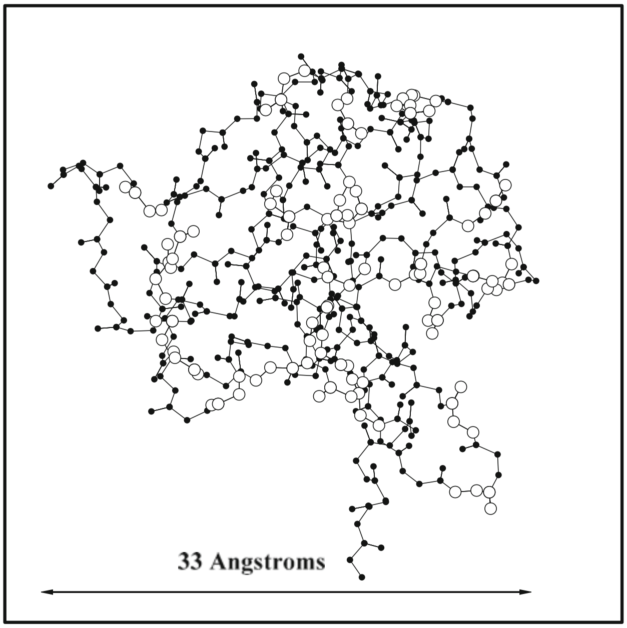

Fig. 7 Statistically-significant effects. *Open circles* mark the neighborhoods where significant (≥ 99% confidence level) MI-changes derived from substitution at *Lys7*, *Arg10*, and *Lys66*. The same type of plot can illustrate the changes in σ_E

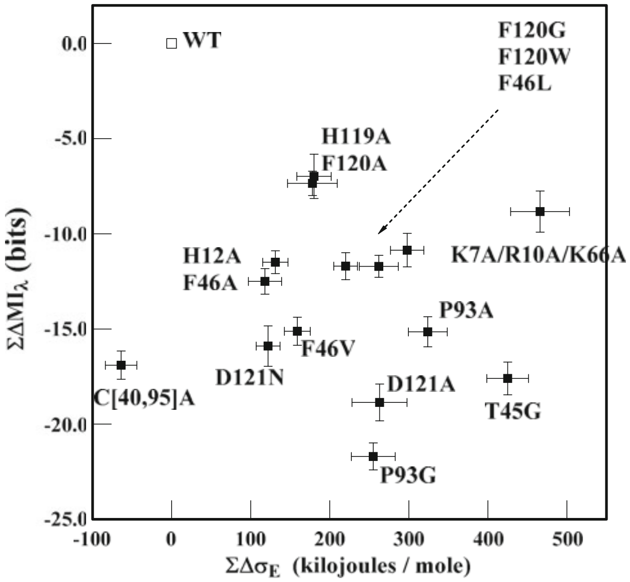

Fig. 8 Folded proteins and state points. The *open square* marks the WT system at 0.00 *bits*, 0.00 kJ/mol. The points for variants are located in reference to the origin. The *error bars* mark the average ± one standard deviation. Each point derives from summing statistically significant differences between WT RNase A and variant molecules

Equation (9) is a straightforward application of the binomial distribution [31]. Accordingly, the probability of obtaining the σ_E results by chance alone is prob(N_+)= 15 / $2^{15} \approx 4.58 \times 10^{-4}$, or in other terms, about 1 in 2200. The analogous calculation applied to the MI results leads to

$\text{prob}(N_-) \approx 3.05 \times 10^{-5}$, or about 1 chance in 33,000. The probability of observing, by chance, state points confined to one quadrant of the Fig. 8 plane is $(4.58 \times 10^{-4}) \times (3.05 \times 10^{-5}) \approx 1.40 \times 10^{-8}$. The authors did not interpret the results to be a mere fluke.

Discussion

This article opened with the condensed primary structure of RNase A. Such a letter sequence is valuable not in itself, but rather its potential. A protein so constructed has the capacity for folding into the conformation shown in Fig. 1. Along the way, the molecule acquires a catalytic function specific to RNA. The importance of the chemistry—and thus the sequence, folding, and variation possibilities—cannot be overstated. Catalyzing RNA hydrolysis is the means for an organism to deactivate genetic material when and where needed [6,15,16,32].

The folding and chemical function are absolutely critical. At the same time, thousands or more variants would be expected to be capable on both accounts. Nature opts for one molecule in an organism by way of the WT. It is important to characterize its signature traits regarding information.

The traits lie in the electronic message sequences ... *(C–C)(C–H)(N–H)*... that can be registered in thermal environments. The WT proves richer in correlations than variants. The wealth has nothing to do with the preponderance of alanine (A) substitutions in the Table 1 library. Tryptophan (W) expresses the greatest correlated information of all the amino acids, non-standard ones included [24]. Yet *F120W* displays a deficit of correlations, and accordingly MI, compared with the WT. In a similar vein, the WT demonstrates lower enthalpy dispersion. This shows that substitutions mimic the effects of crystal impurities [33]. That is to say, substitutions in a native protein and in a crystal reduce the spatial correlations and broaden the energy distribution. While we have found this behavior to be especially pronounced for RNase A, it is not confined to this enzyme. The authors have directed the same analysis to lysozyme molecules, 3D structures of which have also been established experimentally [34]. For a 21-member library, 13 demonstrated significantly lower MI compared with the WT while σ_E was enhanced for 19 of the systems. The probability of observing such MI-results by chance is

$$\text{prob}(N_-) = \frac{\frac{N!}{N_-!(N-N_-)!}}{2^N} = \frac{\frac{21!}{13!(21-13)!}}{2^{21}} \approx 0.0970 \qquad (10)$$

For the σ_E-results, the probability is:

$$\text{prob}(N_+) = \frac{\frac{N!}{N_+!(N-N_+)!}}{2^N} = \frac{\frac{21!}{19!(21-19)!}}{2^{21}} \approx 1.00 \times 10^{-4} \qquad (11)$$

While the MI results for lysozyme are not as dramatic as for RNase A, the σ_E-data are equally compelling. As best we can tell at this writing, lysozyme presents a more complicated system than RNase A, one that warrants additional study. Among other things, the lysozyme backbone is not conserved to the same degree as in RNase A upon amino acid substitution. Alterations of the backbone may provide mechanisms for lysozyme to recover some of the message correlations lost by substitution.

The lessons from ribonuclease can best be explained by viewing the protein primary structure as an encrypted message. While opaque in meaning, the primary structure is the one-dimensional (1D) program that directs 3D folding [35]. The reading and execution of the program obtain from nature expending free energy—that lost by solvent exclusion and the formation of hydrogen, Coulombic, and van der Waals bonds. The 1D program appears to be without rationale. The 3D output makes sense of it, however, by materializing the binding, recognition, and active sites. By transforming the encrypted message, nature purchases a compact, stable molecule with a definitive set of chemical functions.

Encrypted messages are legion in programming and communication. For example, the first paragraph of the second section, in an encrypted form, appears as follows.

Eel r isIdBurnzcdelnecfyneiyms n oarrvetees htimriaaten we tesitutaerrrsho ineeaee spdr nfim ptth-itasor lhnne ege tmahc ifigotoatitlefhrsb rl onwimeot rl sh iiiBe morcrocn hnotalenrws anc.nodcuoni a anIaimlewcnm toap as libolm/trenkexsfctecoraao amtrun vaatsmhinaael lepyttted ihBrntoreoa arn obboawi es annbndiiftfBro oal ninrbrowcam tetycasenideto wodmi anasopirksunil cot tt arapenomgsrtindbio puoacou ctnennteeds iomois rsdtisnmu nhi.Bregn[ni2o seencytt3ca] rntneitac.teEonhottnesnvs efps doiiirbyynrsfis ee ecdsoDcuNf olemte Ass cropcto ro[ed2mapuronotp2 h]lildnieyce.ow mamnc zele pTnlrycomaytholea sent ear ymaynsaacnad certs ndts dhtiasora o mtDpanessnhNhadeAadi ensz.re w a.cx

The above registers zero response in systems such as the reader because the correlations—pair, triplet, and higher order—are absent in the code units.

Rearranging the letters spatially fixes matters. If the order of units is transformed from

1	2	3	4	5	6	7	8	9	10	11	12...
E	e	l	−−	r	−−	i	s	I	d	B	u...

to

1	101	201	301	401	501	601	602	502	402	302	202...
E	n	z	y	m	e	s	−−	w	e	r	e...

the message acquires a bona fide function. The impact on the reader is generated at the same time as the spatial correlations are born. Note that the transformation involves pairing code units that are spatially removed from one another in the source. The correlations render the message not only operational but robust. A reader need not scrutinize every letter to comprehend the message. This is because code units, placed in non-encrypted formats, carry information about their neighbors. This feature—MI—lowers the work required of the message recipient for interpretation.

Now suppose that one or a few units in the encrypted source were altered by arbitrary substitution, e.g.,

Eel r isIdBur*n*zcdelnecfyneiyms → Eel r isIdBur*k*zcdelnec fyneiyms
n oarrvete*e*s htimriaaten → n oarrvete*g*s htimriaaten
tesituta*e*rrrsho ineeaee → tesituta*t*rrrsho ineeaee

The results do not affect the inscrutability of the encrypted source message one way or the other. Yet the substitutions *do* impact the correlations post transformation. The above have the following effects in the translated version:

Eel r isIdBur*k*zcdelnecfyneiyms → →→ E*k*nzymes were. . .
n oarrvete*g*s htimriaaten → →→ Enzym*g*s were. . .
tesituta*t*rrrsho ineeaee → →→ Enzymes w*t*re. . .

Correlations arise by judicious spatial transformation of the code units. Arbitrary replacements tend to eradicate one to several correlations in the end product. Usually the effects are modest, whereby the receiver of the information can still infer the sense of the post-transformation message. Some replacements wield high impact, however. In language, this would be the case if the verb or subject of a sentence were lost.

The results of the project become significant for two reasons. First is that the information of electronic messages offers another structure-based means for contrasting WT and variants. As mentioned, WT RNase A is readily distinguished from variants by bench chemistry. Yet the procedures are lengthy and tedious by and large. Further, enzyme kinetic measurements require precise foreknowledge of the substrate—RNA as opposed to carbohydrates, phospholipids, or other compounds. Brownian computations can be applied to the X-ray or NMR structure data for a protein, whether or not the substrate or chemical function are known. The importance should not be understated. For a folded protein, it is difficult to anticipate the catalytic functionality (hydrolase, oxidoreductase, etc.) and substrate identity on the basis of structure alone. For example, only with PDB access could a chemist pinpoint the correct substrate for the structures of Figs. 1 and 2.

Proteins are not static. They evolve over time in conjunction with the host organism [36,37]. WT RNase A is critical to its host given its activity and specificity toward RNA. Yet there appears little reason why both attributes cannot be enhanced, or that additional chemical functions cannot be incorporated. In other words, the molecule with N- to C-terminal formula

KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNL
TKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQT
NCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHI
IVACEGNPYVPVHFDASV

is not necessarily the best nature can or intends to do.

This leads us to the second reason the results of the project are significant. The WT protein is notable for its correlated information and enthalpy dispersion properties. It would be exceedingly valuable to identify the amino acid substitutions which enhance the ABA correlations and narrow the dispersion even further. The molecules designed from these substitutions could be without peer in degrading genetic material in therapy applications, in addition to providing ancillary chemical functions. Brownian computations are high-throughput because of the simplicity of random walks and record tapes—there are no differentials or integrals to compute in establishing the message correlations. Identifying lead variants by computation appears easier than preparing and evaluating proteins by bench chemistry.

Summary and closing

Proteins present enormous variant possibilities, each representing a modification of Angstrom-scale information. The study for this article focused on the electronic message contrasts of WT RNase A and variants in thermal environments. The native molecule demonstrates greater correlated information and lower enthalpy dispersion in its message space. The results offer another means of discriminating WT and variants based on 3D structure data such as X-ray. Also afforded is a strategy for enhancing the enzyme activity and incorporating new chemical functions by substitutions. At present the authors are quantifying the information properties of inhibitor-bound proteins. RNase A is generous with lessons here as well.

Acknowledgments Discussions with Professors Miguel Ballicora, Duarte Mota De Freitas, Ken Olsen, and Bryan Pickett helped educate the authors about proteins. JLG acknowledges the support of a GAANN Fellowship and Fourth Year Graduate Fellowship from Loyola University Chicago.

References

1. Smyth DG, Stein WH, Moore S (1963) The sequence of amino acid residues in bovine pancreatic ribonuclease: revisions and confirmations. J Biol Chem 238:227–234
2. Chatani E, Hayashi R, Moriyama H, Ueki T (2002) Conformational strictness required for maximum activity and stability of bovine pancreatic ribonuclease A as revealed by crystallographic

study of three Phe120 mutantsat 1.4 Å resolution. Protein Sci 11:72–81. doi:10.1110/ps.31102

3. Park C, Schultz LW, Raines RT (2001) Contribution of the active site histidine residues of Ribonuclease A to nucleic acid binding. Biochemistry 40:4949–4956. doi:10.1021/bi0100182
4. Pearson MA, Karplus PA, Dodge RW, Laity JH, Scheraga HA (1998) Crystal structures of two mutants that have implications for the folding of bovine pancreatic ribonuclease A. Protein Sci 7:1255–1258. doi:10.1002/pro.5560070522
5. Fisher BM, Schultz LW, Raines RT (1998) Coulombic effects of remote subsites on the active site of ribonuclease A. Biochemistry 37:17386–17401. doi:10.1021/bi981369s
6. Raines RT (1998) Ribonuclease A. Chem Rev 98:1045–1066. doi:10.1021/cr960427h
7. Keleman BR, Schultz LW, Sweeney RY, Raines RT (2000) Excavating an active site: the nucleobase specificity of ribonuclease A. Biochemistry 39:14487–14494. doi:1021/bi001862f
8. Hopfinger AJ, Duca JS (2001) Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. J Chem Inf Comput Sci 41:1367–1387. doi:10.1021/ci0100090
9. Godden JW, Stahura FL, Bajorath J (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 40:796–800. doi:10.1021/ci000321u
10. Stahura FL, Godden JW, Xue L, Bajorath J (2000) Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. J Chem Inf Comput Sci 40:1245–1252. doi:10.1021/ci0003303
11. Bajorath J (2000) Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. Mol Divers 5:305–313. doi:10.1023/A:1021321621406
12. Sadowski J, Kubinyl H (1998) A scoring scheme for discriminating between drugs and nondrugs. J Med Chem 41:3325–3329. doi:10.1021/jm9706776
13. Ajay Walters WP, Murcko MA (1998) Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J Med Chem 41:3314–3324. doi:10.1021/jm970666c
14. González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007) A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. J Proteome Res 6:904–908. doi:10.1021/pr060493s
15. Agüero-Chapin G, González-Díaz H, de la Riva G, Rodríguez E, Sánchez-Rodríguez A, Podda G, Vázquez-Padrón RI (2008) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from *Schizosaccharomyces pombe*, prediction, and experimental assay of a new sequence. J Chem Inf Model 48:434–448. doi:10.1021/ci7003225
16. González-Díaz H, Dea-Ayuela MA, Pérez-Montoto LG, Prado-Prado FJ, Agüero-Chapín G, Bolas-Fernández F (2009) QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new *Leishmania infantum* protein. Mol Divers 14:349–369. doi:10.1007/s11030-009-9178-0
17. McGaughey GB, Culberson JC, Feuston BP, Kreatsoulas C, Maiorov V, Shpungin J (2006) Scoring of KDR kinase inhibitors: using interaction energy as a guide for ranking. Mol Divers 10:341–347. doi:10.1007/s11030-006-99037-1
18. Roy DR, Sarkar U, Chataraj PK, Mitra A, Padmanabhan J, Parthasarathi R, Subramanian V, Van Damme S, Bultinck P (2006) Analyzing toxicity through electrophilicity. Mol Divers 10:119–131. doi:10.1007/s11030-005-9009-x
19. Tanabe K, Lucic B, Amic D, Kurita T, Kaihara M, Onodera N, Suzuki T (2010) Prediction of carcinogenicity for diverse chemicals based on substructure grouping and SVM modeling. Mol Divers 14:789–802. doi:10.1007/s11030-010-9232-y
20. Petit B, Potenzone R Jr, Hopfinger AJ, Klopman G, Shapiro M (1979) A hierarchal QSAR molecular structure calculator applied to a carcinogenic nitrosamine data base. In: ACS Symposium Series on computer-assisted drug design, chap 25. pp 553–581. doi:10.1021/bk-1979-0112.ch025
21. Graham DJ, Greminger JL (2009) On the information expressed in enzyme primary structure: lessons from ribonuclease A. Mol Divers 14:673–686. doi:10.1007/s11030-009-9211-3
22. Bennett CH (1982) Thermodynamics of computation—a review. Intl J Theo Phys 21:905–940. doi:10.1007/BF02084158
23. Tribus M, McIrvine EC (1971) Energy and information. Sci Am 225:179–184
24. Graham DJ, Malarkey C, Schulmerich MV (2004) Information content in organic molecules: quantification and statistical structure via Brownian processing. J Chem Inf Comput Sci 44:1601–1611. doi:10.1021/ci0400213
25. Graham DJ, Schulmerich MV (2004) Information content in organic molecules: reaction pathway analysis via Brownian processing. J Chem Inf Comput Sci 44:1612–1622. doi:10.1021/ci040022v
26. Graham DJ (2005) Information content and organic molecules: aggregation states and solvent effects. J Chem Inf Model 45: 1223–1236. doi:10.1021/ci050101m
27. Kadonoso T, Chatani E, Hayashi R, Moriyama H, Ueki T (2003) Minimization of cavity size ensures protein stability and folding: structures of Phe46-replaced bovine pancreatic RNase A. Biochemistry 42:10651–10658. doi:10.1021/bi034499w
28. Schultz LW, Hargraves SR, Klink TA, Raines RT (1998) Structure and stability of the P93G variant of ribonuclease A. Protein Sci 7:1620–1625. doi:10.1002/pro.5560070716
29. Schultz LW, Quirk DJ, Raines RT (1998) His...Asp catalytic dyad of ribonuclease A: structure and function of the wild-type, D121N, and D121A enzymes. Biochemistry 37:8886–8898. doi:10.1021/bi972766q
30. Pauling L (1970) General chemistry, appendix VIII. Dover, New York
31. Kittel C (1986) Elementary statistical physics, chap 6. Dover, New York
32. Dyer KD, Rosenberg HF (2006) The RNase A superfamily: generation of diversity and innate host defense. Mol Divers 10: 585–597. doi:10.1007/s11030-006-9028-2
33. Ashcroft NW, Mermin ND (1976) Solid state physics, chap 30. Holt, Rinehart, and Winston, New York
34. Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR (1965) A three-dimensional Fourier at synthesis 2 Angstrom resolution. Nature 206: 757–761. doi:10.1038/35090602
35. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230. doi:10.1126/science.181.4096.223
36. Ogawa T (2006) Molecular diversity of proteins in biological and defense systems. Mol Divers 10:511–514. doi:10.1007/s11030-006-9048-y
37. Zeldovich KB, Shakhnovich EI (2008) Understanding protein evolution: from protein physics to Darwin selection. Ann Rev Phys Chem 59:105–127. doi:10.1146/annurev.physchem.58.032806.104449