SciForum
Mol2Net

Information Signatures of Viral Proteins: A Study of Influenza A Hemagglutinin and Neuraminidase

Daniel J. Graham[1,*], Samuel Barlow[2], Diego F. Cucalón[3], and Jordan C. Hauck[4]

[1] Department of Chemistry and Biochemistry, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626, USA; dgraha1@luc.edu;
[2] Loyola University Chicago; sbarlow1@luc.edu.
[3] Loyola University Chicago; dcucalon@luc.edu
[4] Loyola University Chicago; jhauck@luc.edu.
***Author to whom correspondence should be addressed; E-Mail: dgraha1@luc.edu; Tel: 1-773-508-3169; Fax: 1-773-508-3086.**

Received: / Accepted: / Published:

Abstract:

Hemagglutinin (HA) and neuraminidase (NA) are glycoproteins encoded by several types of viral particles. Most notably, they exercise complementary chemical functions during infection and propagation of influenza A: infection of a host is initiated by HA while NA catalyzes the release of newly-made viral particles. The antibodies of the molecules form the means of classifying the influenza A subtypes: H1N1, H2N2, H3N2, etc.. Given the risks of viral exposure to global host populations, intense effort is directed toward understanding the molecular mechanisms. Further, the design and formulation of drugs which subvert the mechanisms are on-going challenges. This research focuses on the primary structure information expressed by the two proteins, applying an information theoretic model from previous research. The amino acid sequences for HA and NA such as

MKARLLILLCALSATD…..

MNPNQKIITIGSICMAI……

are parsed for their correlated information, both the total accumulation and fluctuations. Data for the HA and NA of multiple influenza A subtypes are illustrated via information signatures and phase plots. This enables sharp contrasts to be drawn between seasonal infectious proteins and ones with high pandemic potential. Overall, the analysis illuminates new ways of evaluating HA and NA molecules for their subtype and virulence based on information properties. Just as important, the results point to mutation strategies for re-directing and attenuating the protein functions.

Keywords: proteins; information; influenza; viruses; hemagglutinin; neuraminidase

1. Introduction

Hemagglutinin (HA) and neuraminidase (NA) are glycoproteins in the surface membrane of influenza particles [1]. Infection of a host is initiated by HA while NA catalyzes the release of newly-made viral particles [2]. The antibodies of the molecules form the means of classifying the influenza A subtypes: H1N1, H2N2, H3N2, etc. [3]. At present, there are at least 16 and 9 known subtypes for HA and NA, respectively. Given the risks of viral exposure to global populations, intense effort is directed toward understanding the molecular mechanisms. Further, the design and formulation of drugs which subvert the mechanisms are on-going challenges [4].

Influenza HA and NA have presented thousands of variants. For example, two HA sequences are:

MKARLLILLCALSATDADTICIGYHANNSTDTVDTVLEKNVTVTH
SVNLLEDSHNGKLCRLKGIAPLQLGKCNIAGWILGNPECESLLSNR
SWSYIAETPNSENGTCYPGDFADYEELREQLSSVSSFERFEIFPKER
SWPKHNITRGVTAACSHAKKSSFYKNLLWLTEANGSYPNLSKSY
VNNKEKEVLVLWGVHHPSNIEDQRTLYRKENAYVSVVSSNYNRR
FTPEIAERPKVRGQAGRMNYYWTLLEPGDKIIFEANGNLIAPWYA
FALSRGLGSGIITSNASMDECDTKCQTPQGAINSSLPFQNIHPVTIG
ECPKYVRSTKLRMVTGLRNIPSIQSRGLFGAIAGFIEGGWTGMVD
GWYGYHHQNEQGSGYAADQKSTQNAINGITNKVNSVIEKMNTQF
TAVGKEFNKLEKRMENLNKKVDDGFLDIWTYNAELLVLLENERT
LDFHDSNVKNLYEKVKNQLRNNAKEIGNGCFEFYHKCDNECMES
VKNGTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASS
LVLLVSLGAISFWMCSNGSLQCRICI Seq. (1)

MEARLLVLLCAFAATNADTICIGYHANNSTDTVDTVLEKNVTVT
HSVNLLEDSHNGKLCKLKGIAPLQLGKCNIAGWLLGNPECDLLLT
ASSWSYIVETSNSENGTCYPGDFIDYEELREQLSSVSSFEKFEIFPKT
SSWPNHETTKGVTAACSYAGASSFYRNLLWLTKKGSSYPKLSKS
YVNNKGKEVLVLWGVHHPPTGTDQQSLYQNADAYVSVGSSKYN
RRFTPEIAARPKVRDQAGRMNYYWTLLEPGDTITFEATGNLIAPW
YAFALNRGSGSGIITSDAPVHDCNTKCQTPHGAINSSLPFQNIHPVT
IGECPKYVRSTKLRMATGLRNIPSIQSRGLFGAIAGFIEGGWTGMI
DGWYGYHHQNEQGSGYAADQKSTQNAIDGITNKVNSVIEKMNT
QFTAVGKEFNNLERRIENLNKKVDDGFLDIWTYNAELLVLLENER
TLDFHDSNVRNLYEKVKSQLKNNAKEIGNGCFEFYHKCDDACME
SVRNGTYDYPKYSEESKLNREEIDGVKLESMGVYQILAIYSTVASS
LVLLVSLGAISFWMCSNGSLQCRICI Seq. (2)

Two NA sequences are:

MNPNQKIITIGSICMAIGTISLILQIGNIISIWVSHSIQTGSQNHTGICN
QRIITYENNTWVNQTYVNISNTNVVAGKDTTSMILAGNSSLCPIRG
WAIYSKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDK
HSNGTVKDRSPYRALMSCPIGEAPSPYNSRFESVAWSASACHDGM
GWLTIGISGPDDGAVAVLKYNGIITEIIKSWRKQILRTQESECVCVN
GSCFTIMTDGPSDGPASYRIFKIEKGKITKSIELDAPNSHYEECSCYP
DTGKVMCVCRDNWHGSNRPWVSFNQNLDYQIGYICSGVFGDNP
RPKDGKGSCDPVNVDGADGVKGFSYRYGNGVWIGRTKSNSSRK
GFEMIWDPNGWTDTDGNFLVKQDVVAMTDWSGYSGSFVQHPEL
TGLDCMRPCFWVELIRGRPREKTTIWTSGSSISFCGVNSDTVNWS
WPDGAELPFTIDK Seq. (3)

MNPNQKIITIGSICMVVGIISLILQIGNIISIWVSHSIQTGNQNHPETC
NQSIITYENNTWVNQTYVNISNTNVVAGQDATSVILTGNSSLCPIS
GWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLND
KHSNGTVKDRSPYRTLMSCPVGEAPSPYNSRFESVAWSASACHD
GMGWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECA
CVNGSCFTIMTDGPSNGQASYKILKIEKGKVTKSIELNAPNYHYEE
CSCYPDTGKVMCVCRDNWHGSNRPWVSFDQNLDYQIGYICSGVF
GDNPRPNDGTGSCGPVSSNGANGIKGFSFRYDNGVWIGRTKSTSS
RSGFEMIWDPNGWTETDSSFSVRQDIVAITDWSGYSGSFVQHPEL
TGLDCMRPCFWVELIRGQPKENTIWTSGSSISFCGVNSDTVGWSW
PDGAELPFSIDK Seq. (4)

The sequences offer detailed information. Yet a computer-unassisted reading of them is bewildering. This is apparent because, among other things, one cannot distinguish the extraordinary from ordinary. The above include formulae allied with the "Spanish flu" pandemic of 1918 [5]. But which ones are these? The correct answers are Seqs. (2) and (4). The reader's uncertainty is understandable given the lengths and complexities of the sequences.

Our approach to proteins has looked for guidance from information theory [6 - 10]. Here we focus on the HA and NA primary structure information. The results draw contrasts between seasonal molecules and ones with high virulence potential. The data further point to mutation strategies for re-directing and attenuating the functions.

2. Proteins and Sequence Information

The approach builds on research from the mid-2000s. Work in this lab quantified the correlated information *CI* expressed by the naturally occurring amino acids based on their atom and covalent bond structure [6, 8]. An average < *CI* > and standard deviation σ_{CI} were established and a dimensionless quantity $Z_{CI}^{(i)}$ was based on each amino acid's *CI* contribution relative to the average *CI*, e.g.

$$Z_{CI}^{(W)} = +2.63 \qquad Z_{CI}^{(F)} = +0.691$$

$$Z_{CI}^{(M)} = -0.128$$

$$Z_{CI}^{(A)} = -0.476$$

There are twenty amino acids and thus sixteen more $Z_{CI}^{(i)}$ to note as in reference [6]. The superscript symbols refer to the amino acid while the numerical value represents the *CI* distance from the average in standard deviation (σ_{CI}) units. The sign reflects whether the amino acid contributes information above or below the natural average. The *Z*-terms largely follow chemical intuition. Tryptophan (W) features a network of aromatic bonds and functional groups; it exerts nearly +3σ_{CI} impact in a protein. Alanine (A) is a simple aliphatic and contributes *CI* below average at ca. -0.5σ_{CI}. The methodology originated in an information study of ribonuclease A and lysozyme [8, 9].

A dimensionless function *G*(*k*) is constructed for a sequence by adding $Z^{(i)}_{CI}$ in the N- to C-terminal order; *k* is a counting index less than or equal to *N* number of residues in the protein. *G*(*k*) tracks the accumulation and fluctuations of information:

$$G(k) = Z^{(i)}_{CI,1} + Z^{(i)}_{CI,2} + Z^{(i)}_{CI,3} + \dots + Z^{(i)}_{CI,k} = \sum_{j=1}^{k \le N} Z^{(i)}_{CI,j} \qquad (1)$$

Proteins generally host a majority of low information residues. As a consequence, *G*(*k*) scales linearly with negative slope and is well accommodating of least squares analysis. The analysis establishes an ensemble of linear regression functions $L_j(k)$ with typical correlation coefficient $R^2 > 0.95$.

The ensemble leads to information signatures { $H_j(k)$ }:

$$\{ H_j(k) \} = \{ G(k) - L_j(k) \} \qquad (2)$$

As examples, *G*(*k*), {$H_j(k)$}for Seqs. (1) and (4) appear in **Figure 1**. The proteins originate from a human host H1N1 isolate harvested in Albany, New York in 1951. The accession details are: gb:CY021821|gi:145279077|UniProtKB:A4U7A6|.

Graphs such as in **Figure 1** serve as signatures of the primary structure information. They reflect more than a molecule's local composition. If a substitution is made at site *j*, the collection in Eq. (2) is altered. The amplitude is impacted at all sites *k* = 1, 2, …, *N*. The information signatures can be strikingly different, depending on the subtype. There are as many signatures as there are HA and NA variants.

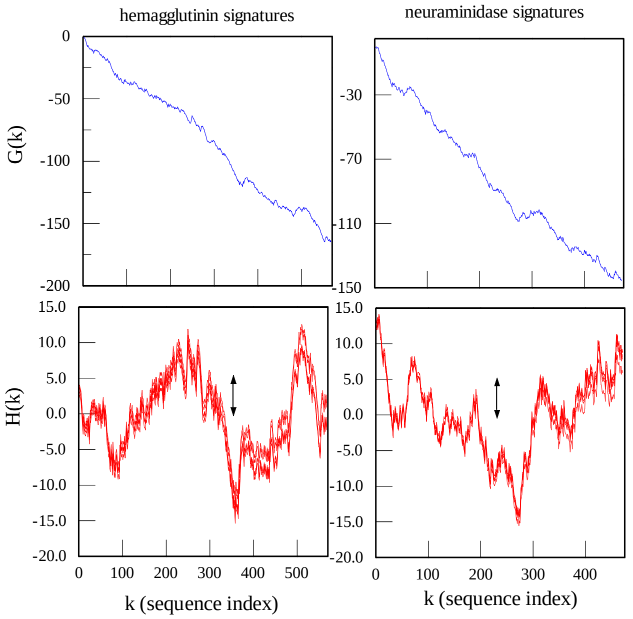

Figure 1: Information Functions for HA and NA. Plots of *G*(*k*) and *H*(*k*) derive from Seqs. (1) and (3) of the **Introduction**. The vertical arrows in the lower panels indicate the standard deviation in the fluctuations of *H*(*k*).

In thermodynamics, the variance of an extensive property such as enthalpy and entropy scales with a capacity [11]. In the same way, the variance in $H_j(k)$ can be viewed in terms of a protein's functional capacity. Molecules composed of only one type of amino acid, e.g. AAAAAAAAA…., offer zero capacity. They are of no biochemical utility because they lack diversity of information. This is borne out in the signatures: their *G*(*k*) trace perfect lines (R^2 = 1.000); corresponding *H*(*k*) express zero amplitude.

The information signature variance is calculated as follows:

$$\sigma_H^2 = \langle H^2 \rangle - \langle H \rangle^2 \qquad (3)$$

The square root σ_H is the standard deviation, so indicated in **Figure 1** by the vertical black arrows. Linear regression computes an ensemble of *H*(*k*) functions from *G*(*k*), and accordingly, a distribution of σ_H. Capacities form robust descriptors of thermodynamic systems; σ_H play equally vital roles regarding protein information.

3. Results

Nearly sixty thousand primary structures were analyzed; the HA and NA sequences were obtained as FASTA downloads from the Influenza Research Database (IRD). The data were catalogued according to viral subtype: H1N1, H2N2, H3N2, and so forth. For every molecule, $G(k)$, $\{ H_j(k) \}$were established along with distributions of the variance and standard deviation. A viral isolate locates a coordinate (plus error bars) on a σ_{HA}, σ_{NA} phase plot. Multiple variants establish a neighborhood of points for a subtype. The distance from one neighborhood to another is dictated by the information effects of antigenic drift and genome re-assortments.

The phase plot for influenza A is illustrated in **Figure 2**. Each filled circle marks the average σ_{HA}, σ_{NA} for the labeled subtype while the bars mark the standard deviations about the averages. Several bars are of widths less than the symbols. These reflect that a sparse number of isolates was available for analysis. That being said, every attempt was made to be exhaustive. **Figure 2** derives from HA and NA across a spectrum of hosts: human, avian, equine, bat, etc.. There will be more neighborhoods to map as new subtypes and hosts are discovered.

Figure 2 teaches two things, the first concerning boundaries. There are astronomical possible variants of HA and NA. The ones selected for viral infection and propagation are concentrated in the following ranges:

$$2.5 \le \sigma_{NA} \le 6.0$$

$$3.8 \le \sigma_{HA} \le 8.0$$

Note the ranges to be substantive despite the intense selection pressure to preserve the protein functions.

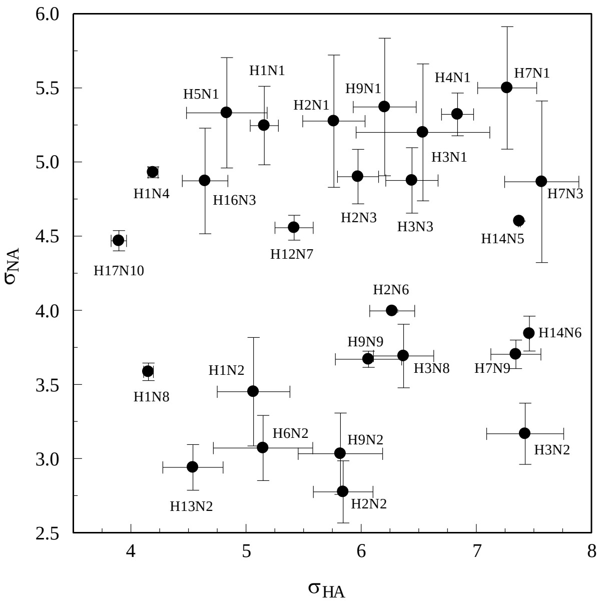

Figure 2: Phase Plot for Influenza HA and NA. Each filled circle marks the average σ_{HA}, σ_{NA} for the labeled subtype. The error bars mark the standard deviations about the averages.

The second lesson is that the neighborhood distribution is markedly uneven. A significant fraction of influenza subtypes clusters in the upper third of **Figure 2** while fewer ones occupy the lower third. Further, there are several low-density regions: these correspond to HA, NA variants which have yet to manifest, or are outright avoided by natural selection.

Information signatures discriminate the subtypes. What do things look like for proteins specific to human populations? For humans, the major circulating strains of influenza A have been H1N1, H2N2, and H3N2; the global pandemic of 1918 was attributed to the first of these [5]. The avian strain H5N1 has rarely infected humans, although it poses high virulence potential.

Figure 3 shows a phase plot based on human host isolates. Different color symbols distinguish the subtypes while the isolate years are included. Not all years are represented as the analysis was directed to complete genomes. The point locus for the 1918 pandemic year sample (Brevig Mission, >gb:AF250356|gi:8572169|UniProtKB:Q9IGQ6|) is marked in red. It is considerably removed from the H1N1 neighborhood. Its nearest neighbors derive from H5N1 isolates.

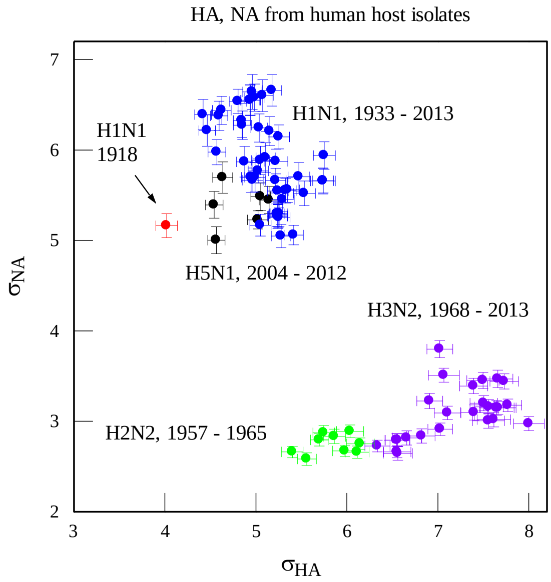

Figure 3: Phase Plot for Human Host HA and NA. The blue symbols are placed by H1N1 samples collected between 1933 and 2013. The red symbol is placed by HA and NA from pandemic year 1918. The black, green, and violet symbols are placed, respectively, by H5N1, H2N2, and H3N2 samples.

4. Discussion

HA and NA exercise complementary functions: the former enables attachment of influenza particles to a cell surface while the latter catalyzes the release [1, 2]. Given the essentialness of the functions, there is significant pressure for variants to manifest over time. Variants enable the virus to sidestep host immune responses and to thwart drug therapy. There are >10^5 HA and NA sequences on record, yet this is a paltry number compared to the possibilities.

Thermodynamic analysis of a system commences with variables and functions of state. This has been the approach to HA and NA in constructing information *G* and *H*. The functions track the information accumulation and fluctuation in a manner dependent on *all* the amino acids. No one site or region is viewed as more important than others.

One learns several things, the first being an information method of evaluating HA, NA pairs. All sequences are confounding by their complexity. However, using a spreadsheet and $Z_{CI}^{(i)}$ look-up table, it is straightforward to compute *G*-functions plus regression *L* and signature *H*. The signatures lead immediately to σ_{HA}, σ_{NA} and the neighborhood of the phase diagram. To be sure, the evaluation is not without tentativeness given the overlap of neighborhoods. However, the analysis points to the information and virulence similarities of the viral subtypes.

.

The second insight is the contrast between seasonal- and pandemic-year proteins. HA and NA from the 1918 pandemic places an outlier point on the phase plot. This placement stems from a lower fluctuation amplitude, compared with that of seasonal proteins. This suggests that lower chemical noise in the primary structures underpins a more invasive chemical function.

The third insight is a strategy for re-directing—and possibly attenuating—the functions. Natural selection favors molecules which promote viral infection and suppresses variants that serve otherwise. We conjecture that the latter type place state points in the low density regions of **Figure 2**.

Figure 4 revisits **Figure 2** and includes four pathways. Each tracks a succession of substitutions on Seqs. (1) and (4). With each substitution, there is a displacement of the σ_{HA}, σ_{NA} coordinate. As the primary structures belong to the H1N1 subtype, each pathway commences near the center of the H1N1 neighborhood and terminates in a zero-to-low density region. The paths are annotated by the following ordered- pair sequences:

Pathways 1	**2**	**3**	**4**	**HA, NA mutations**
E199G, D199V	P504A, H144K	Y209N,C279A	E51R, T362I	
F432V, D329C	Y246K, V291N	L335T, D79W	R238K, T381G	
R514I, C238F	P135Y, G342R	S179K, A271K	G411V, R52W	
S275Y, S166P	A302S, K369R	K328N, L127H	K511Q, I20S	
E260D, P326D	W553M, N235T	A379H, M188C	L547C, K150F	
E449G, P337T	W553K, G333H	V428I, W458E	D456W, V149W	
L250F, H126A	I530E, Y208L	R344Q, S168H	G76W, L22Y	
N177T, H185M	L512A, P328K	N448D, V75K	Q386I, G109T	
L417M, N141P	S160N, N171Q	L37M, P169R	A156D, N325I	

The pathways (and countless more) are readily charted using a forced random walk algorithm. One selects a target locale on the phase plot. The sequences are then subject to trial substitutions. With each trial, *G*, *H*, and σ_{HA}, σ_{NA} are computed. The substitutions are accepted if the state point is inched closer to the target and declined otherwise. Each pathway in **Figure 4** is traversed via nine pair-substitutions. This demonstrates that the proteins do not have to be radically altered for the information signatures to move out of the virulent neighborhood of origin. In **Figure 4**, pathways 1 and 2 direct HA and NA away from *all* the subtype neighborhoods. In contrast, pathways 3 and 4 cross territory allied with highly virulent subtypes. In re-directing HA and NA functions, the upward-going pathways 1 and 2 would seem preferable. Molecules with information removed from the active neighborhoods would likely offer diminished potency, yet stimulate some production of host antibodies. This would enhance the overall immunity of host populations.

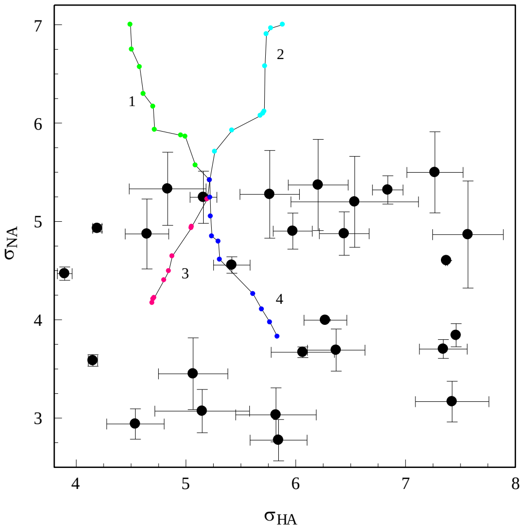

Figure 4: Pathways for Re-locating Phase Points. Each pathway tracks a succession of amino acid substitutions on the H1N1 Seqs. (1) and (3) of the **Introduction**. The pathways are annotated above.

——

Summary and Closing

The primary structure information expressed in influenza HA and NA was investigated using a model established in previous research. The model enabled computation of signatures based on the accumulation and fluctuation of information. The signatures were encapsulated in *G*, *H*-functions and phase plots. These illuminated information methods for discriminating variants and attenuating the molecular functions.

Acknowledgements

The authors appreciate the input of Donald May about information theory in reference to biological systems. The authors equally appreciate the insights of Miguel Ballicora about natural selection and viral systems. The authors thanks Loyola University Chicago for support of this research.

References

1. Levine AJ (1992) Viruses, Scientific American Library, Chapter 8.
2. Voyles BA (1993) The Biology of Viruses, Mosby-Year Book, St. Louis.
3. Kawaoka Y, Neumann G in Influenza Virus: Methods and Protocols (2012), Kawaoka Y, Neumann G eds, Humana Press, New York, Chapter 1.
4. Roberts NA (2001) Prog. Drug Res. 56:195-237.
5. Kolata GB (1999) Flu: the Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It, Farrar, Straus, and Giroux, New York.
6. Graham DJ, Malarkey C, Schulmerich MV (2004) J Chem Info Comp Sci, 1601.
7. Graham DJ (2013) Prot J 32:275-287. 10.1007/s10930-013-9485-2.

8. Graham DJ, Greminger JL (2009) Mol Divers 10.1007/s11030-009-9211-3.

9. Graham DJ, Greminger JL (2011) Mol Divers 10.1007/s11030-011-9307-4.

10. Graham DJ, May D, Grzetic S, Zumpf J (2012) Prot J 31:550-563. 10.1007/s10930-012- 9432-7.

11. Goodstein DL (1985) States of Matter, Dover, New York, Chapter 1.

Author Contributions

The authors contributed jointly to the paper regarding protein sequence retrieval, computer programming,, information data analysis, and biological significance.

Conflicts of Interest

The authors declare no conflict of interest.