Global informatics and physical property selection in protein sequences

Harold A. Scheraga[a,1] and S. Rackovsky[a,b,1]

[a]Department of Chemistry and Chemical Biology, Baker Laboratory, Cornell University, Ithaca, NY 14853; and [b]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Contributed by Harold A. Scheraga, January 4, 2016 (sent for review December 10, 2015; reviewed by Robert L. Jernigan and Jeffrey Skolnick)

The degree of informatic independence between the physical properties of amino acids as encoded in actual protein sequences is calculated. It is shown that no physical property can be identified that carries significantly less information than others and that the information overlap between different properties and different length scales along the sequence is essentially zero. These observations suggest that bioinformatic models based on arbitrarily selected sets of physical properties are inherently deficient.

protein bioinformatics | physical properties | information theory | Fourier analysis

Protein bioinformatics originated with the use of alignment-based methods to compare protein sequences. In recent years, there has been a great increase in the use of "knowledge-based" methods, in which sequences are characterized by assigning to their component amino acids numerical values of physical properties that are believed to be important. These properties are usually selected to provide a quantitative basis for an intuitive picture of the physical chemistry of amino acids. Subsequent analysis is carried out using either a detailed description of the sequences of interest, which involves a consideration of local sequence characteristics, or a set of sequence-averaged property values, which involves discussion of global sequence characteristics (1, 2).

In previous work, we have considered various informatic aspects of the sequence–structure relationship in proteins. It was shown that there is structural uncertainty encoded in both local sequence information (3) and the global properties of sequences (4). These uncertainties constitute intrinsic limitations of knowledge-based protein bioinformatics. In this work, we examine investigator-imposed limitations resulting from the selection of "representative" sets of physical properties to characterize the amino acids. We ask whether an arbitrarily preselected set of amino acid physical properties can act as proxies for others that are not included. We also ask whether physical properties can be identified that are significantly less important than others and therefore, candidates for exclusion from knowledge-based models of the sequence–structure relationship. The following results are shown.

i) The information–theoretic relationship between different global sequence physical properties can be computed.
ii) Sequence-averaged and -specific variables can be treated in a unified manner.
iii) When physical properties are expressed using an exhaustive and nonredundant representation, their longitudinal sequence distributions are found to be unrelated informatically. This observation implies that the use of a limited set of preselected variables is guaranteed to result in information loss.
iv) The informatic contributions of all variables are found to be of the same order of magnitude. This fact implies that there are no variables that are safely negligible.

Results

Our approach is based on two tools. The first tool is the physical description of amino acids using the Property Factor Representation developed by Kidera et al. (5, 6). This representation is based on a factor analysis of all of the available physical properties of 20 amino acids, and it was shown that all of these data (comprising 189 separate property sets) can be represented by 10 property factors, which together, carry 86% of the variance of the entire dataset. Therefore, the physical properties of an amino acid X can be represented numerically by a 10-vector **X**:

$$\mathbf{X} = \left(f_X^{(1)}, f_X^{(2)}, \ldots, f_X^{(10)}\right), \quad [1]$$

where $f_X^{(i)}$ is the ith property factor of amino acid X. The central point for this work is that, by construction, the property factors are complete and orthonormal, and therefore, **X** is an excellent numerical representation of the totality of physical properties of the amino acid.

The second tool is the representation of protein sequences in Fourier space. We have discussed details of this approach extensively and showed its utility in bioinformatic studies (4, 7–11). In this approach, the sequence of a protein is written in terms of the 10 vectors embodied in Eq. **1**. This process gives a representation of the N-residue sequence in terms of 10 N-member numerical strings, each of which describes the course of one property factor from N terminus to C terminus. Each of these strings is Fourier-transformed, giving a set of (sine and cosine) Fourier coefficients, which are labeled by two indices: k, the wave number, and l, an index in the range $1 \leq l \leq 10$ that identifies the property factor string that has been transformed. Several facts are of importance in this connection.

i) Each Fourier coefficient contains information from the entire sequence of the protein.
ii) The $k = 0$ (cosine) Fourier coefficients are sequence averages of 10 property factors and contain no information about the actual longitudinal arrangement of amino acids along the chain.

Significance

Many bioinformatic investigations of protein sequence–structure relationships are based on preselected sets of amino acid physical properties. We investigate the extent to which such preselection is justified on information–theoretical grounds. It is shown that neither sequence-dependent nor -averaged properties can be identified as informatically negligible and that no physical properties can be identified that carry sufficient information about other variables to act as surrogates. This result implies that knowledge-based studies of proteins must be based on complete, nonredundant physical property sets.

Author contributions: H.A.S. and S.R. designed research; S.R. performed research; S.R. analyzed data; and H.A.S. and S.R. wrote the paper.

Reviewers: R.L.J., Iowa State University; and J.S., Georgia Tech.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: has5@cornell.edu or srr87@cornell.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1525745113

Table 1. Diagonal sine (SIN) elements of the mutual information matrix

SIN, SIN	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 9$	$l = 10$
$k = 1$	2.10	2.13	2.04	1.95	2.12	2.08	2.10	2.12	2.13	2.12
$k = 2$	2.12	2.15	2.06	1.98	2.12	2.10	2.12	2.12	2.13	2.12
$k = 3$	2.14	2.13	2.05	1.99	2.13	2.10	2.11	2.11	2.12	2.13
$k = 4$	2.15	2.12	2.07	2.01	2.12	2.11	2.12	2.11	2.13	2.13
$k = 5$	2.16	2.14	2.10	2.03	2.13	2.11	2.13	2.12	2.13	2.12
$k = 6$	2.16	2.12	2.13	2.06	2.14	2.12	2.12	2.10	2.14	2.12
$k = 7$	2.15	2.12	2.14	2.09	2.13	2.11	2.14	2.10	2.12	2.12

iii) A Fourier coefficient with wave number $k > 0$ encodes characteristics of the chain that occur at the spatial length scale N/k. Therefore, coefficients with low k values describe the organization of the sequence on relatively long scales, and those with high k describe the sequence in terms of local characteristics.

iv) We have shown (11) that sequence differences between families of proteins of different structures are encoded in a limited number of low k Fourier coefficients. We find that coefficients with wave numbers in the range $0 \leq k \leq 7$ are important for encoding global folding information.

It should be emphasized that, as a corollary of points *ii* and *iii*, sequence-specific (longitudinal) and sequence-averaged information on physical properties are treated on an equal footing by this approach and can be compared systematically and rigorously.

The architecturally relevant range of Fourier coefficients, thus, includes those with $0 \leq k \leq 7$ for all values of l, the physical property index. We ask whether any subset of this complete set of Fourier coefficients carries the information necessary to characterize a set of sequences. For this hypothesis to be true, it must be shown that some Fourier coefficients encode information that is also present in others. The appropriate tool for investigating this question is mutual information, an information–theoretic function that measures the degree to which two random variables are independent. The mutual information of variables X and Y is given by

$$I(X,Y) = \sum_i \sum_j p(x_i)p(y_j)\log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right). \qquad [2]$$

Here, $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the probability distributions of the individual variables. It can be seen that, if X and Y are independent, so that $p(x,y) = p(x)p(y)$, then $I(X,Y) = 0$. Note that $I(X,X) = H(X)$, where $H(X)$ is the entropy of the random variable X.

In this work, the random variables of interest are centered, normalized Fourier coefficients (12). We calculate the complete set of $I(Z_k^{(l)},Z_{k'}^{(l')})$, where $Z_k^{(l)}$ is either sine or cosine Fourier coefficient (as defined in *Methods*). Each value represents the mutual information over a very large set of protein sequences (described in *Methods*) with low pairwise sequence identity. One can think of these values as elements of a matrix. From the foregoing discussion, it can be seen that the off-diagonal elements measure the degree to which the information encoded in the two different Fourier coefficients overlaps. The diagonal elements give the width of the distribution of the Fourier coefficient in question. This width is a measure of the amount of information encoded in that coefficient.

We find that there is a large difference between the magnitudes of diagonal and off-diagonal elements of this mutual information matrix. The off-diagonal elements, for $0 \leq k \leq 7$ and $1 \leq l \leq 10$, and either sine or cosine Fourier coefficients have values $I < 10^{-1}$ nats. In contrast, diagonal elements for $k \neq 0$ have values of $I \geq 1.95$ nats and for $k = 0$, $I > 1.4$ nats. The off-diagonal elements are essentially zero, indicating that no Fourier coefficient encodes information that is also encoded in a coefficient with different k or l.

Discussion and Conclusions

Because the Fourier coefficients encode, to a good approximation, both the complete physical properties of the amino acids and the detailed longitudinal sequence information that distinguishes between folds, it will be seen, from the negligible values of the off-diagonal mutual information, that neither physical property nor relevant degree of sequence resolution can be omitted from a knowledge-based calculation without incurring an informatic penalty.

Examination of the diagonal elements of the mutual information matrix (Tables 1, 2, and 3) provides additional insight into this phenomenon. The entropy of a distribution measures the width of that distribution and quantitates the uncertainty in the value of the variable that the distribution represents. A variable can be omitted from a knowledge-based calculation only if the uncertainty in its value is small enough that no significant variation will be overlooked through its omission. The entropies of the diagonal elements for $k > 0$, however, all have very similar values (Tables 1 and 2). The entropies of the $k = 0$ diagonal elements are slightly smaller than those for $k > 0$ but also, all have roughly the same size (Table 3). The variation in an omitted variable will, therefore, be of the same magnitude as the variations of those included. Hence, no case can be made for the omission of any property value or any length scale within the significant range in considering the properties of protein sequences.

Unlike the intrinsic uncertainties in protein informatics identified previously (3, 4), the investigator-dependent uncertainties that we have addressed herein can be avoided by using globally oriented, statistically complete tools of the kind discussed above and in *Methods*.

Table 2. Diagonal cosine (COS) elements of the mutual information matrix

COS, COS	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 9$	$l = 10$
$k = 1$	2.14	2.18	2.05	1.98	2.14	2.08	2.10	2.11	2.15	2.14
$k = 2$	2.16	2.17	2.06	1.99	2.14	2.10	2.11	2.12	2.16	2.12
$k = 3$	2.15	2.13	2.05	1.99	2.13	2.10	2.12	2.12	2.14	2.12
$k = 4$	2.16	2.14	2.08	2.01	2.13	2.11	2.14	2.12	2.13	2.11
$k = 5$	2.15	2.13	2.10	2.03	2.13	2.13	2.11	2.11	2.13	2.13
$k = 6$	2.17	2.14	2.11	2.08	2.13	2.12	2.13	2.13	2.13	2.12
$k = 7$	2.15	2.13	2.13	2.08	2.14	2.12	2.13	2.12	2.13	2.11

Table 3. Diagonal $k = 0$ cosine elements of the mutual information matrix

	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 9$	$l = 10$
COS ($k = 0$), COS ($k = 0$)	1.85	1.78	1.56	1.44	1.55	1.58	1.51	1.61	1.43	1.55

Methods

The details of the Fourier approach have been discussed extensively in previous work (7, 8, 11). We provide here details of the calculation described above.

The random variables of interest are the centered, normalized Fourier coefficients for a protein sequence given by

$$Z_k^{(l)} = \frac{c_k^{(l)} - \left\langle c_k^{(l)} \right\rangle_N}{\sigma\left(c_k^{(l)}\right)}, \quad [3]$$

where $c_k^{(l)}$ is an unnormalized (sine or cosine) Fourier coefficient with wave number k arising from the lth property factor. The angle brackets denote an average over all possible permutations of the original, WT N-residue sequence, and σ is the associated SD. [We have shown (8) that the latter statistical quantities can be calculated analytically.] The effect of this normalization is to remove any dependency on sequence composition alone, so that the random variable explicitly encodes information about the specific linear arrangement of amino acids in the sequence. The $k = 0$ (cosine) Fourier coefficient depends only on sequence composition, because it represents an average of the lth physical property over the sequence in question. It is not normalized, because both the average and SD are, by definition, zero.

In this work, we use a protein dataset based on the CATH sequence/structure database (ref. 13; www.cathdb.info). This dataset, which we have used in previous studies (4), contains 12,011 domains drawn from the CathDomainSeqs.S60.ATOM.v.3.2.0 dataset. The sequences in this set have no more than 60% sequence identity.

To calculate the mutual information (Eq. **2**), we need the individual distributions of all Fourier coefficients [$p(x)$] as well as the joint distributions for all pairs of coefficients [$p(x,y)$]. It was convenient to use 21-bin [and (21×21) bin] histograms for this purpose. This subdivision provided a reasonable compromise between resolution and sparseness. The ranges of the (dimensionless) variables observed for the proteins in our database were $-5.0 < c_k^{(l)} < 5.0$ for $k \neq 0$ and $-1.1 < c_0^{(l)} < 1.0$ for all l.

ACKNOWLEDGMENTS. This research was supported by NIH Grant GM-14312 and National Science Foundation Grant MCB-10-19767.

1. Saxena A, Sangwan RS, Mishra S (2013) Fundamentals of homology modeling steps and comparison among important bioinformatics tools: An overview. *Science Int* 1(7): 237–252.
2. Ruiz-Blanco YB, Paz W, Green J, Marrero-Ponce Y (2015) ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics* 16(2015):162.
3. Rackovsky S (1993) On the nature of the protein folding code. *Proc Natl Acad Sci USA* 90(2):644–648.
4. Rackovsky S (2015) Nonlinearities in protein space limit the utility of informatics in protein biophysics. *Proteins* 83(11):1923–1928.
5. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* 4(1): 23–55.
6. Kidera A, Konishi Y, Ooi T, Scheraga HA (1985) Relation between sequence similarity and structural similarity in proteins: Role of important properties of amino acids. *J Protein Chem* 4(5):265–297.
7. Rackovsky S (1998) "Hidden" sequence periodicities and protein architecture. *Proc Natl Acad Sci USA* 95(15):8580–8584.
8. Rackovsky S (2006) Characterization of architecture signals in proteins. *J Phys Chem B* 110(38):18771–18778.
9. Rackovsky S (2009) Sequence physical properties encode the global organization of protein structure space. *Proc Natl Acad Sci USA* 106(34):14345–14348.
10. Rackovsky S (2010) Global characteristics of protein sequences and their implications. *Proc Natl Acad Sci USA* 107(19):8623–8626.
11. Rackovsky S (2013) Sequence determinants of protein architecture. *Proteins* 81(10): 1681–1685.
12. Scheraga HA, Rackovsky S (2014) Homolog detection using global sequence properties suggests an alternate view of structural encoding in protein sequences. *Proc Natl Acad Sci USA* 111(14):5225–5229.
13. Sillitoe I, et al. (2015) CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res*, 10.1093/nar/gku947.