SHORT COMMUNICATION

Property-Based Sequence Representations Do Not Adequately Encode Local Protein Folding Information

A.D. Solis and S. Rackovsky*

Department of Pharmacology and Biological Chemistry, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, New York 10029

ABSTRACT **We examine the informatic characteristics of amino acid representations based on physical properties. We demonstrate that sequences rewritten using contracted alphabets based on physical properties do not encode local folding information well. The best four-character alphabet can only encode ~57% of the maximum possible amount of structural information. This result suggests that property-based representations that operate on a local length scale are not likely to be useful in homology searches and fold-recognition exercises. Proteins 2007;67:785–788. © 2007 Wiley-Liss, Inc.**

Key words: bioinformatics; reduced alphabets; homology search; fold recognition; amino acids

INTRODUCTION

Reduced amino acid alphabets are an important tool in protein bioinformatics.[1] They appear both explicitly in folding and simulation studies, and implicitly in homology searches, in which groups of amino acids are declared equivalent for the purpose of sequence matching, in order to determine structural similarity. Homology searching, as commonly practiced, is based on an evolutionary model.[2–4] The probability of an amino acid replacement is calculated from data that reflect evolutionary processes, using alignments of sequences which are judged to be related and quantitating the likelihood of specific replacements. Recently, however, it has become clear that it is not necessary for two sequences to exhibit similarity in any conventionally accepted sense, in order for their structures to be similar. This is one of the deepest observations in current protein science. It means that we do not understand the manner in which structure is encoded in sequence, and that evolutionary models are not adequate to define structural similarity. A new approach to the definition of sequence similarity, more closely related to the physics of structure determination, would be a significant advance in homology studies. It would seem natural to compare sequences without evolutionary bias using reduced alphabets, based on amino acid physical properties, which must ultimately determine native structure. Appropriately chosen matches based on property similarity might be expected to select for local folding similarity. A number of alphabets have, in fact, been constructed based on various physical properties.[1]

In previous work,[5,6] we have examined the question of reduced alphabet construction from a different viewpoint. We began from a consideration of structure, and asked whether a reduced amino acid alphabet can be constructed which maximally retains structural information. It was shown that, for any given alphabet size, an optimal alphabet can indeed be constructed, and the amount of structural information which it encodes can be computed. It has been shown[1] that the resulting alphabet performs well in sequence alignment and fold-assessment tests.

The optimal alphabet does not correspond to those which have been proposed based on property considerations. It is therefore appropriate to ask whether property-based representations of protein sequence are capable of encoding sufficient structural information to make them useful in structural homology searches. It is this point which we investigate in this work. We show that property-based sequence alphabets do not encode sufficient local folding information to be useful in structural homology searches. This observation implies that the information encoded by physical properties is distributed over the entire sequence, rather than acting locally. On a practical note, it suggests that homology searches using physical

Grant sponsor: National Library of Medicine, National Institutes of Health; Grant number: 2R01 LM-06789.

*Correspondence to: S. Rackovsky, Department of Pharmacology and Biological Chemistry, Mount Sinai School of Medicine, Box 1603, One Gustave L Levy Place, New York, NY 10029. E-mail: shelly@camelot.mssm.edu

Received 30 August 2006; Revised 27 November 2006; Accepted 19 January 2007

Published online 23 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21434

properties must be carried out using nonlocal representations.

METHODS

The strategy we adopt is straightforward.[5] The first step is the selection of a representation of amino acids in terms of their physical properties. The amino acids are then clustered into a predefined number of groups, and the sequences of a protein database are rewritten in terms of the resulting reduced alphabet. The distributions of structures associated with all N-grams in the database are constructed, and the total of their entropies is calculated. This total entropy is used to calculate the information gain associated with the reduced alphabet, relative to the structure distribution in the absence of sequence information. This quantity measures the degree to which the reduced alphabet in question encodes local structural information.

The first matter which must be addressed is the choice of a property-based representation of amino acids. Many property sets have been defined for the 20 naturally occurring amino acids. A set of "physically reasonable" properties manually chosen to represent the amino acids will, in general, have two problems: The representation will be incomplete, and there will be significant correlation between parameters representing the various properties. These difficulties have been treated by Kidera et al.,[7,8] who carried out a factor analysis of a comprehensive, redundant set of amino acid properties, and showed that the data can be represented by a set of 10 factors, which together carry 86% of the variance of the entire data set.* Therefore, to an excellent approximation, an amino acid **X** can be represented numerically by a 10-vector of properties:

$$\mathrm{X} = (f_x^{[1]}, f_x^{[2]}, \ldots, f_x^{[10]}) \qquad (1)$$

A more general version of this representation can be constructed by attaching a weight to each factor:

$$\mathrm{X} = (\mu_1 f_x^{[1]}, \mu_2 f_x^{[2]}, \ldots, \mu_{10} f_x^{[10]}) \qquad (2)$$

By varying the components of the weight set {μ| i = 1,2,. . .,10} we can generate representations that adjust the relative importance of the properties arbitrarily. Because a factor analysis results in an essentially orthonormal basis set for the data on which it is performed, adjusting the 10 weights in (2) gives a good approximation to all possible property-based sequence representations. Each reweighting of the representation gives a new reduced alphabet, and a new value of the information gain.

Amino Acid Clustering

Given a specified weight set {μ}, corresponding to a representation of an amino acid X in vector form, the 20 amino acids are clustered as follows. A distance function between amino acids i and k is given by

$$d_{ik} = \left[\sum_{j=1}^{10} (\mu_j f_i^{[j]} - \mu_j f_k^{[j]})^2\right]^{1/2} \qquad (3)$$

We write the objective function which gives a figure of merit for a particular clustering in the following form:

$$D = \sum_{r} \sum_{i,k \subset r} d_{ik} \qquad (4)$$

where r is a cluster number. The best clustering of the amino acids is that which minimizes D. This is readily found by a simple Monte Carlo search of cluster space. The number of clusters is prespecified. We study the four-cluster case in this work.

Mutual Information

We wish to compare the entropy (essentially the width) of the distribution of structures associated with a particular reduced alphabet to that of the distribution of structures in the absence of sequence labeling. This is formally expressed as the information gain:

$$I(c, s^{\alpha}) = H(c) - H(c|s^{\alpha}) \qquad (5)$$

where c is an index specifying the local backbone conformation, and s^{α} specifies the sequence written in the reduced alphabet a. The structural entropy $H(c)$ is given by

$$H(c) = \sum_{n} p(c_n) \ln[p(c_n)] \qquad (6)$$

where $p(c)$ is the fractional occurrence of conformation n in the database. The conformations are described in this work using the GBM representation, based on a discretization of the virtual bond backbone.[5] The corresponding equation for the entropy of the distribution with sequence specified is

$$H(c|s^{\alpha}) = \sum_{m} \left\{ \sum_{n} p(c_n|s_m^{\alpha}) \ln[p(c_n|s_m^{\alpha})] \right\} p(s_m^{\alpha}) \qquad (7)$$

where s_m^{α} denotes the sequence fragment m, written in the reduced alphabet a, and $p(s_m^{\alpha})$ is the fractional occurrence of s_m^{α} in the database.

* The first four factors, which together carry a significant part of the total variance, correspond essentially to a single property. The remaining six are linear superpositions of several properties and are labeled by the most important contributing property. The ten properties are as follows:

1. Helix/bend preference
2. Side-chain size
3. Extended structure preference
4. Hydrophobicity
5. Double-bend preference
6. Amino acid composition
7. Flat extended preference
8. Occurrence in alpha region
9. pK
10. Surrounding hydrophobicity

TABLE I. Results From Property-Based Clustering

Scheme	Factor WTS 1234567890	Amino Acid ACDEFGHIKLMNPQESTVWY	$H(c\mid s)$ (nat)	$I(c,s)$ (cnat)
	INFO OPTD[a]	12112312122141111222	3.7888	16.00
I	1111111111	23441142334113322241	3.9085	4.03
	1000000000	24421311122331444143	3.8875	6.13
	0100000000	44231432133323142211	3.9276	2.12
	0010000000	41442223222241313341	3.9250	2.38
	0001000000	41231424314243322414	3.9189	2.99
	0000100000	34331124242122341131	3.9182	3.06
	0000010000	21442214331424424242	3.9454	0.34
	0000001000	42311234221412231344	3.9495	−0.07
	0000000100	43441144434123322431	3.8993	4.95
	0000000010	12232421314221133341	3.8955	5.33
	0000000001	23122134131424321121	3.9467	0.21
	1001001000	21432142322413343211	3.8739	7.49
II	2012000100	34134234344123111422	3.8652	8.36
III	3003000221	13213423133241122344	3.8577	9.11
IV	3004000222	13213423133241122344	3.8577	9.11
	4014001131	13213423133241122344	3.8577	9.11
	4114011430	13213423133241122344	3.8577	9.11
	4104001220	13213423133241122344	3.8577	9.11
V	Many 5’s[b]	13213423133241122344	3.8577	9.11

[a] The amino acid grouping shown in previous work[5,6] to give the maximum information gain.
[b] 291 unique factor weight sets out of a total of 27,884 local minima reached in the Monte Carlo search.

The Database

We used the PDBselect database of Hobohm and Sander (http://bioinfo.tg.fhgiessen.de/pdbselect/) to construct the nonredundant data set. We selected the high-resolution structures of 1036 protein chains whose sequences share no more than 25% (pairwise) identity. The resulting data set has, in total, 210,995 amino acid residues.

In this work, we consider four-letter alphabets. We also consider a number of resolutions for searching the weight set $\{\mu\}$. With no loss of generality, we can allow each weight to vary over the range $0 \leq \mu_i \leq 1$. We discretize this range in order to search the weight space, allow the 10 weights to vary independently, and calculate the information gain arising from each of the resulting weight sets. Five discretization schemes were used:

I. [0.0, 1.0]
II. [0.0, 0.5, 1.0]
III. [0.0, 0.33, 0.67, 1.0]
IV. [0.0, 0.25, 0.5, 0.75, 1.0]
V. [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]

Using the first four schemes, it is possible to exhaustively search the weight grid. An exhaustive search of the grid is not possible using Scheme V, and therefore a Monte Carlo maximization of the information gain is carried out in this case. The reliability of the discretization concept will be reflected in the convergence rate of information gain as the number of discrete weight values is increased.

RESULTS AND DISCUSSION

In Table I we show the results of the calculation. For each weighting scheme, we show the actual factor weights (renomalized to integers for convenience), the cluster composition, the entropy of the resulting structure distribution, and the information gain resulting from the choice of reduced alphabet. We observe the following points:

- The structurally optimized four-letter alphabet gives an information gain of 16.00 cnats.
- Within scheme I, we give results for three cases—that in which all properties are weighted equally [corresponding to Eq. (1)], that in which only one property is weighted and all others have zero weight (all 10 possibilities), and the optimal weighting.[5] It will be seen that the best clustering arising from equal weighting of all the properties encodes 4.03 cnats—25% of the maximum possible. The best of the single-property weightings, in which factor 1 (helix-bend preference) is weighted, gives an alphabet which encodes 6.13 cnats of structural information. The optimal scheme I weighting, in which factors 1, 4, and 7 (helix-bend preference, hydrophobicity and flat extended preference) are weighted, encodes 7.49 cnats of structural information.
- The best scheme II weighting, in which factors 1 and 4 are again predominant, encodes 8.36 cnats of structural information.
- The best scheme III weighting again shows a prominent role for factors 1 and 4, and encodes 9.11 cnats of information.
- There are four scheme IV weightings that lead to the same optimal clustering, and once again indicate a central role for factors 1 and 4. The information gain is again 9.11 cnats.
- Although we were unable to exhaustively search the weight grid in scheme V, our Monte Carlo optimization

finds extensive degeneracy of optimal weighting sets, all of which lead to the same clustering of amino acids into a reduced alphabet. Once again we find that the information gain arising from this clustering is 9.11 cnats.

The latter three observations indicate that the gridding scheme for searching weight space has indeed converged, and that 9.11 cnats is the maximum structural information which one can encode in a property-based reduced alphabet. This is only 57% of the maximum possible structural information encodable in a four-letter reduced amino acid alphabet.

It is important to translate this information–theoretic result into practical terms. The use of physical properties as labels for the 20 amino acids will result in the matching of sequence fragments whose local structural properties are significantly dissimilar and in misassignment of local folding. The use of property-based reduced alphabets is therefore not likely to be helpful in homology searching, fold detection, structure prediction, or the modeling of folding processes. Since the native structure of a protein must, however, be determined by physical properties, we conclude that, in order to be useful, property-based sequence representations must be designed to reflect long-range, rather than purely local characteristics of the sequence. We are currently investigating this idea.

REFERENCES

1. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins: Struct Funct Bioinf 2006;63:986–995.
2. Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Methods Enzymol 1983;91:524–545.
3. Henikoff S, Henikoff J. Amino Acid substitution matrices from protein blocks. Proc Nat Acad Sci USA 1992;89:10915–10919.
4. Godzik A. Fold recognition methods. Methods Biochem Anal 2003;44:523–544.
5. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins: Struct Funct Genet 2000;38:149–164.
6. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins: Struct Funct Genet 2002;48:463–486.
7. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Prot Chem 1985;4:23–55.
8. Kidera A, Konishi Y, Ooi T, Scheraga HA. Relation between sequence similarity and structural similarity in proteins: role of important properties of amino acids. J Prot Chem 1985;4:265–297.