

Information Theory exercise: Amino Acid Pair Contact Probabilities

Vincent A. Voelz

E-mail: vvoelz@gmail.com

Math Bio Boot Camp 2006

University of California at San Francisco, San Francisco, CA 94143

August 29, 2006

1 Introduction

Statistical potentials have long been used in successful approaches toward protein structure prediction. In particular, pairwise contact potentials have been thought to be useful in capturing some of the essential features of protein folding, such as the hydrophobic effect, ion/salt pairing, disulfide pairing, etc. But how much information do these pairwise statistical potentials contain? This is the question posed by Cline et al. [1] in a recent paper, in which they calculate a **mutual information** measure to characterize the usefulness of pairwise contact potentials.

We will roughly follow the aims Cline et al. paper, and try to calculate some information-theoretic measures on a set of data compiled from protein pairwise contacts.

2 Data set

A set of 3465 proteins were randomly chosen from the PDB (one from each SCOP class), and from this were calculated the frequencies of a contact being formed between amino acid X and amino acid Y . Two amino acids were considered to be in contact if their C_α atoms were closer than 6.5 Å.

The raw data (in `pairprobs.dat`) consists of a (symmetric) matrix of counts $N(X, Y)$ for all amino acid pairs, where

$$X, Y \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

The columns (and rows) of this matrix are arranged in the above order.

3 Problems

1. First, normalize the data to obtain $p(X, Y)$, and sum rows (or columns) to get the marginal distributions $p(X)$ and $p(Y)$ (they should be the same).
2. Calculate the correlation $p(X, Y) - p(X) - p(Y)$. Which contacts are over-represented? Which contacts are under-represented? Does your result correspond to your biophysical intuition?
3. Calculate the uncertainty of the marginal distributions, $H(X)$ and $H(Y)$ (they should be the same).
4. Calculate the uncertainty of the joint distribution, $H(X, Y)$.
5. Calculate the mutual information of the joint distribution, $M(X, Y) = H(X) + H(Y) - H(X, Y)$. How many bits of mutual information are there? Do you think this is a significant number?
6. If you feel daring, try to repeat all of the above with the following variants of simple amino acid alphabets:

TABLE I. Alphabets Used for Relating Contact Patterns to Amino Acid Attributes

Alphabet	Contents	
I	Cys and other	
II	Positive, negative, other	
III	Hydrophobic and polar	
IV	Cys, positive, negative, other polar, other hydrophobic	
V	Standard alphabet of amino acid types	

Alphabet	Category	Contents
I	Cys	Cys
	Other	All others
II	Positive	Arg, His, Lys
	Negative	Asp, Glu
	Neutral	All others
III	Hydrophobic	Ala, Cys, Gly, Ile, Met, Phe, Pro, Trp, Tyr, Val
	Polar	Arg, Asn, Asp, Glu, Gln, His, Lys, Ser, Thr
IV	Cys	Cys
	Positive	Arg, His, Lys
	Negative	Asp, Glu
	Other polar	Asn, Gln, Ser, Thr
	Other hydrophobic	Ala, Gly, Ile, Leu, Met, Phe, Pro, Tyr, Val

Figure 1: Five alternative simpler amino acid alphabets used in the Cline et al. paper.

References

- [1] Melissa S. Cline, Kevin Karplus, Richard H. Lathrop, Temple F. Smith, Jr. Robert G. Rogers, and David Haussler. Information-theoretic dissection of pairwise contact potentials. *PROTEINS: Structure, Function and, Genetics*, 49:7–14, 2002.