Diversity of antibody receptor proteins in the B cell repertoire

Thierry Mora[1], Aleksandra M. Walczak[1,2], William Bialek[1,2] and Curtis Callan[1,2]

Short Abstract — **Recent experiments that provide a nearly exhaustive sampling of B cell antibody sequences in a single individual allow for the statistical characterization of the immune repertoire. The observed variability of the B cell antibody sequences can be explained neither by their genomic origin nor by a model of independent residues. To describe the diversity of the B cell antibody repertoire, we construct models of the sequences based on the principle of maximum entropy. Although the model relies only on pairwise correlations between amino acids at different positions, it agrees with local and global properties of the data. We show that correlations between residues greatly reduce diversity with respect to a model where amino acids are independent of each other. Further, the proposed model unveils a rugged landscape in the distribution of antibody sequences, suggesting possible signatures of adaptation to antigenic challenges.**

Keywords — **B cells, immune repertoire, maximum entropy models, antibody receptor proteins.**

I. Background

Biology offers many examples of families of proteins that perform similar yet slightly different functions. Characterizing the amino acid sequences that make up a given family constitutes an important challenge. Previous studies [1] have focused on classes of related proteins coming from different species or involved in different biological pathways, but that carry out the same basic function. Here we study a class of proteins, the B-cell antibody receptors, which play a key role in the immune system [2]. These proteins show great diversity, even within a single individual, and each of them fulfills a slightly different function. Furthermore, the repertoire of these proteins changes as the immune system adapts to pathogen infections. Because this adaptation occurs on short time scales (compared to Darwinian evolution), this class of proteins provides an ideal system in which to investigate protein diversity and the relation between sequence and function through selective pressure.

We have analyzed recently published data obtained by Weinstein et al. [3], which consists of a nearly exhaustive list of the RNA sequences making up the B-cell antibody repertoires of 14 single zebrafish individuals.

II. Model

We restrict our study to a subregion of receptor sequences, called the D region, which is involved in antigen binding, and is the most variable and the most adaptive part of the sequence. We consider D sequences from a single zebrafish as being drawn from a probability distribution, which we model by a translation invariant maximum entropy model constrained by pairwise correlations between amino acids [4]. Specifically, we look for a model distribution that maximizes Shannon's entropy while matching the pairwise amino acid frequencies measured in the data, and likewise for the distribution of sequence lengths. The model fit is performed by a combination of Monte Carlo sampling and gradient descent.

III. Results

Our model correctly reproduces local as well as global, emergent properties of the sequence ensemble, including triplet amino acid frequencies, which are not fitted by the model. Correlations are shown to dramatically decrease diversity (measured by entropy) with respect to an independent model. In agreement with data but in disagreement with an independent model, we predict that the distribution of antibodies follows Zipf's law (which states that the frequency of an antibody sequence is inversely proportional to its rank). We use the model to quantify repertoire diversity and specificity across individuals. The model assigns an effective energy to each sequence. We explore the structure of the corresponding energy landscape, and find possible signatures of adaptation to antigenic challenges in the presence of many metastable states.

References

[1] M Socolich, SW Lockless, WP Russ, H Lee, KH Gardner, Ranganathan (2005) Evolutionary information for specifying a protein fold. *Nature* **437**, 512; WP Russ, DM Lowery, P Mishra, MB Yaffe, R Ranganathan, (2005) Natural–like function in artificial WW domains. *Nature* **437**, 579; M Weigt, RA White, H Szurmant, JA Hoch, T Hwa (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **106**: 67.

[2] KM Murphy, P Travers, M Walport (2007) *Immunobiology: The Immune System (Janeway)*, (2007), Garland Science; 7 edition.

[3] JA Weinstein, N Jiang, RA White, DS Fisher, SR Quake (2009) High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science* **324**:807.

Acknowledgements: This work was supported in part by NSF Grant PHY–0650617 and by NIH Grant P50 GM071598; T. M. was supported by the International Human Frontiers Science Program Organisation.

[1]Joseph Henry Laboratories of Physics and Lewis-Sigler Institute for Integrative Genomics. Princeton University, Princeton NJ 08544. E-mail: tmora@princeton.edu

[2]Princeton Center for Theoretical Science, Princeton University, Princeton NJ 08544.