Data mining of protein families using common peptides.

Assaf Gottlieb, Uri Weingart and David Horn[§]
School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel.

Abstract
Predicting the function of a protein from its sequence is typically addressed using sequence-similarity. Here we propose a motif-based approach, using supervised motif extraction from protein sequences belonging to one functional family. The resulting deterministic motifs form Common Peptides (CPs) that characterize this family, allow for data mining of its proteins and facilitate further partition of the family into clusters.

Introduction
Kunik, et al., (2007) have recently introduced a novel method for classifying enzymes based on Specific Peptides (SPs) that are strings of amino-acids, extracted from enzyme sequences using the motif extraction algorithm MEX (Solan, et al., 2005). Motif extraction was carried out in an unsupervised fashion, and SPs were selected from the resulting motifs according to their specificity to levels of the Enzyme Commission (EC) 4-level functional hierarchy.
We propose to apply MEX in a supervised fashion on a family of proteins, which may be a family of enzymes belonging to one EC number, but can be any other protein family. After some further processing (selection of length > 4 amino-acids and elimination of degeneracy) we define a set of Common Peptides (CPs) characterizing the protein family. We develop a search methodology that allows us, on the basis of the number of CP hits on a protein sequence, to decide if the protein belongs to the same family.

Results

ThyX is one of two thymidylate synthase families. We have analyzed data (Stern, et al., 2008) containing 136 thyX sequences from different organisms. Applying MEX to the data we extracted 168 distinct peptides. They are found to cover 133 sequences, i.e. they occur at least once on 98% of the data.
We have studied the occurrence of CPs (number of hits) on enzyme sequences of the training set, and compared it to the occurrence of the same CPs on unrelated enzymes. Since CPs have not been selected according to specificity to a particular EC number, they may be found on sequences of enzymes whose function is unrelated to that of the family from which they were extracted. Nonetheless the occurrence distribution, as shown in Figure 1, is very different. In unrelated enzymes one may see one CP hit, and rarely two. These numbers are consistent with a background random model. This is also true when searching for hits on the other family of thymidylate synthase, thyA. Thus, although the CPs have not been selected to be specific to thyX, the occurrence of several of them on a sequence is quite specific. Moreover, applying biclustering to the matrix of sequences *vs.* CPs, we obtain clusters that are consistent with grouping the relevant species according to evolutionary proximity.

The DNA gyrase subunit B enzyme, gyrB, has been proposed by Watanabe, et al. (2001) for the purpose of spanning a database for identification and classification of bacteria. GyrB in a single copy protein, and is one of several protein families belonging to EC 5.99.1.3. Analyzing 100 gyrB proteins from the ENZYME database we obtain a set of CPs that is applied to the Sargasso-Sea metagenomic data, estimating the number of distinct gyrB genes by requiring the occurrence of at least two consistent matches out of this CP pool. We find that 989 out of over 1 million proteins can be identified as such. This result is quite close to the number of maximal fragment depth of 924 quoted in Table 2 of Venter et al. (2004).

Olfactory receptors form a large family, belonging to the superfamily of G-protein coupled receptors (GPCR). We have used 3733 OR sequences (Gottlieb et al., 2008) representing the OR repertoires of 8 vertebrates extracting a pool of CPs. A background random model predicts that the appearance of three or more peptides occurs with a false-positive probability of about 1%. Fig. 2 implies that there is some real overlap between olfactory CPs and general GPCR ones, nonetheless the separation of the two distributions is evident.

REFERENCES

Gottlieb, A, Olender, T, Lancet, D, and Horn, D. (2008) *in preparation*.

Kunik, V, Meroz, Y, Solan, Z, Sandbank, B, Weingart, U, Ruppin, E, and Horn, D.(2007) . Functional representation of enzymes by specific peptides. *PLOS Comp. Biol*. 3(8):e167

Lagerström, M.C., A.R. Hellström, D.E. Gloriam, T.P. Larsson, H.B. Schiöth, and R. Fredriksson. 2006. The G Protein–Coupled Receptor Subset of the Chicken Genome. *PLoS Comput Biol* **2:** e54.

Solan Z, Horn D, Ruppin E, Edelman S. Unsupervised learning of natural languages. Proc. Natl. Acad. Sci. USA 102: 11629-11634 (2005).

Stern, A. *et al.* (2008) On the evolution of thymidine synthesis: a tale of two enzymes and a virus., *submitted for publication*

Venter, J. C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D. et al. (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304. 66 – 74.

Watanabe, K., Nelson, J., Harayama, S. and Kasai, H. (2001). ICB database: the gyrB database for identification and classification of bacteria.

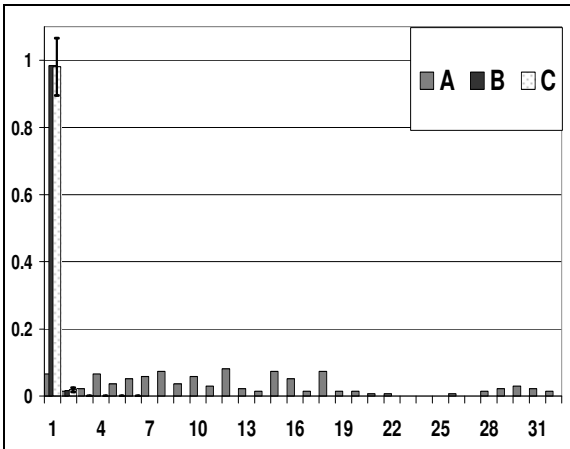

Fig. 1. Number of CPs observed on each of the thyX sequences (A), is compared with the observation on sequences from all other enzymes (B), and with that of a random model (C). All three cases are normalized to total area = 1.

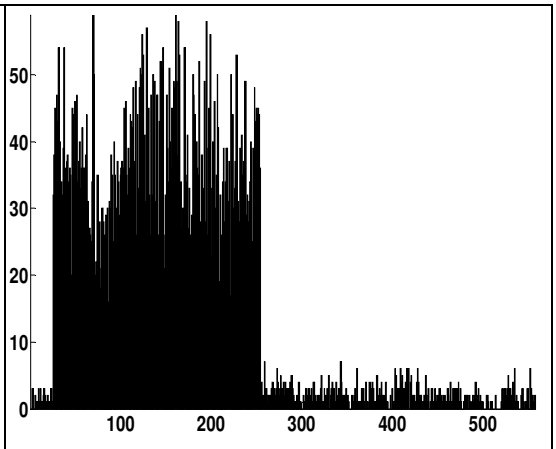

Fig. 2. Histogram of the number of peptides hits for each chicken GPCR from (Lagerström et al., 2006). Only the area of high values from x=26 to x=254 contains intact and pseudogenes of chicken ORs.

§ to whom correspondence should be addressed