BST281: Genomic Data Manipulation, Spring 2018

Wednesday 12: Proteins: Proteomics, Families, and Folds

**Proteomics**

* The direct detection and quantification of the proteins in a biological system
	+ Can also assay protein “states” [post-translational modifications (PTMs), e.g. phosphorylation]
	+ Provides high-confidence detection of proteins/validation of putative coding genes
	+ Provides more accurate protein abundance estimate than inference via mRNA abundance
* Antibody arrays use protein-specific antibodies to detect and quantify proteins (~microarray)
	+ Useful for small sets of specific proteins (e.g. cytokines)
* 2D gels separate a protein mixture based on two properties, typically size and pH/charge
	+ Compare 2D coordinates to reference positions for identification
* Tandem mass-spectrometry (MS-MS) = shotgun sequencing for proteins
	+ Protein mixture is stratified (e.g. by liquid chromatography) to reduce complexity
	+ Proteins are digested (e.g. by Trypsin) into short peptide fragments
	+ Peptide fragments are isolated in a first round of MS
	+ Fragments are shattered into ions along peptide bonds
	+ “Stepped” mass/charge (m/z) ratios reveal amino acid sequence during second round of MS

**Protein families**

* Protein sequences map to functions/folds many-to-one (genotype-phenotype mapping)
* Multiple strategies for defining protein “families” (sequences compatible with a given function)
* Global protein similarity
	+ Clusters of Orthologous Groups (COGS) – Reciprocal best BLAST hit relationships
	+ UniRef – Clustering of the protein universe at different levels of sequence identity
* Local protein similarity (domains/motifs)
	+ Pfam (and related approaches) define local multiple sequence alignments (MSAs) for domains
	+ Multiple domains can occur in a protein = domain architecture

**Protein structure determination and data**

* Methods for determining protein structure
	+ X-ray crystallography (gold standard, requires protein to crystallize)
	+ Nuclear Magnetic Resonance (NMR; can probe dynamic processes, but limited to small proteins)
	+ *Ab initio* folding on a computer based on physical models (hard)
	+ Knowledge-based folding (homology modeling, threading, fragment assembly)
* Structure data (x, y, z coordinates of atoms) are stored in the Protein Data Bank
* Structures are organized into hierarchies based on the similarity of their folds (e.g. CATH, SCOP)
	+ Structures at top levels in the hierarchy may not be obviously related by descent, or at all

**Hidden Markov Models (HMMs) – part of slide “extras”**

* A Markov model is a set of states and transition probabilities
* In a HMM, we don’t observe states directly, but rather “clues” emitted in the form of symbols
	+ Ex. Genomic CpG dinucleotide symbols emitted more often in “gene” state than “not gene” state
* Protein family MSAs are represented as HMMs to enable highly sensitive detection in new sequences
	+ Each MSA column is a state; emission probabilities are amino acid frequencies in that column
	+ Additional states allow deletions and insertions

**Suggested reading**

* Pevsner: Ch. 12, pp. 539-575 (Proteomics); Ch. 13, pp. 589-625 (Protein structure)
* Bloom, Jesse D., et al. "[Structural determinants of the rate of protein evolution in yeast](https://academic.oup.com/mbe/article/23/9/1751/1014274/Structural-Determinants-of-the-Rate-of-Protein)." *Molecular biology and evolution* 23.9 (2006): 1751-1761.