Computational biology in the 21st century: Algorithms that scale Supplement

1 Gene expression

While sequence is important, scientists are also interested in measuring gene expression, or the quantity of a particular gene product (usually a protein or more recently RNA) present in the cell. Gene expression data is useful for relating genotype to phenotype, and is important in studying many diseases. Usually obtained through microarrays or RNA-seq technologies, expression data is quantitative, as each gene from a sample is associated with a numeric expression level, and high-dimensional, as many thousands of genes may be analyzed at a time. Resources such as NCBIs Gene Expression Omnibus (GEO), which pool together many diverse expression studies, now allow researchers to analyze thousands of samples generated by others to obtain meaningful biological insights. More importantly, many of these biological insights can only be gleaned when viewed in the context of hundreds or thousands of gene expression samples [1].

2 Protein structure

Protein structure data, primarily determined through X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, describes the coordinates of each atom, in Angstroms, of a protein in three-dimensional space. This information is important to understanding how a protein functions, because proteins interact physically with other molecules and are geometrically constrained. Protein structure data are commonly stored in the Protein Data

Bank (PDB); each PDB entry has an associated resolution indicating the accuracy of the structural description. Thus a PDB entry at its most basic level contains a list of atoms, along with which amino acid they belong to and their spatial coordinates as floating-point numbers.

Over the course of evolutionary time, structure is known to be more highly conserved than sequence [5], which is to say that it does not change as rapidly. When the structure of a protein is known but its biological function or evolutionary relationships are not, researchers may search for structurally similar proteins that are better studied [3]. Classical tools for this involve performing pairwise structural alignments to look for geometric similarity; DALI [4] is still widely used, along with other aligners such as FATCAT [7] and Matt [6]. Due to the complexity of protein structures, these programs generally take significant amounts of time, especially to align multiple structures. For example, the DALI webserver can take as much as an hour to return results for a single query. FragBag [2] accelerates protein structure search by approximating structural alignments by instead comparing the 'bag-of-words' from each structure. Analogous to a term-frequency vector in information retrieval, this bag-of-words indicates the abundance of particular, short structural motifs within a protein.

References

[1] B. Berger, J. Peng, and M. Singh. Computational solutions for omics data. *Nature Reviews Genetics*, 14(5):333–346, 2013.

[2] I. Budowski-Tal, Y. Nov, and R. Kolodny. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences*, 107(8):3481–3486, 2010.

[3] J.-F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Current opinion in structural biology*, 6(3):377–385, 1996.

[4] L. Holm and C. Sander. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, 20(11):478–480, 1995.

[5] K. Illergård, D. H. Ardell, and A. Elofsson. Structure is three to ten times more conserved than sequencea study of structural response in protein

cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

[6] M. Menke, B. Berger, and L. Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology*, 4(1):e10, 2008.

[7] Y. Ye and A. Godzik. Fatcat: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research*, 32(suppl 2):W582–W585, 2004.