CS 626 - COMPUTATIONAL MOLECULAR BIOLOGY
Chapter 1

(A) Sequence annotation

a. Elements of DNA and protein sequences

Proteins, DNA (Deoxyribonucleic acid) and RNA (Ribonucleic acid), are all linear polymers and (arguably) the most important biological molecules. Linear polymers in general are made of monomers chemically linked in a one-dimensional sequence. They may adopt a well-defined three-dimensional structure, as proteins do, or exist in multitude of alternative conformations (e.g. polymers of hydrocarbons). For DNA and RNA there are only four types of monomers, which we denote by the characters: (A, C, G, T) or (A, G, T, U) respectively. A, C, G, and T are used to create especially long DNA molecules that store the genetic code. Proteins are made of twenty types of monomers (amino acids) supporting the greater diversity of shapes and functions that are found in proteins and are more difficult to obtain using four types of monomers. Of course, RNA, which also acts as an enzyme, suggests that four types only can be pushed quite far.

Segments of the DNA (genes) code proteins. Therefore, a prime target of sequencing projects is the accurate determination of coding regions. The role of the non-coding parts is still not clear, and it suggests an intriguing research topic that is *not* the focus of the present book. Even within a gene that codes one protein we are not free of surprises (e.g. more non-coding sequences). Consider for example the phenomenon of introns and exons. Non-coding segments (introns) separate coding parts (exons) even of a single protein chain. Coding parts could be merged in more that one way (alternative splicing), leading to alternative proteins that were built from the same “chunks” of DNA.

The coding part of the DNA sequence determines protein sequences. The genetic code (the translation from DNA bases (monomers) to amino acids) is given in table 1.

Table 1

Three bases (XYZ) code each amino acid. The bases at the left are at the first position (X), the four columns in the middle stand for the second position (Y) and the third base (Z) is at the last column. Note the high degeneracy of the genetic code, and the multiplicity of many amino acids including the STOP signal. The third base is highly degenerate. Tryptophan (Trp) and Methionine are exceptions. Note also that the degeneracy at the codon level does not imply the same degeneracy at the level of amino acids. Arginine is highly popular at the codon level (six repeats). It is not as frequent in actual protein.

1^{st} position	2^{nd} position	2^{nd} position	2^{nd} position	2^{nd} position	3^{rd} position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

A list of the twenty amino acids in alphabetic order, their symbols and their properties is given in table 2. The chemical and physical properties of the amino acids are roughly characterized in the last column. “Hydrophobic” mean an amino acid that is not well solvated in water. These amino acids tend to aggregate at the core of the protein avoiding contact with the water solution. Polar and charged residues are much better solvated and usually found at surface of the compact proteins hape.

Alanine	Ala	A	Hydrophobic
Cysteine	Cys	C	Hydrophobic / sulfur bridges
Aspartic acid	Asp	D	Charged (negative)
Glutamic acid	Glu	E	Charged (negative)
Phenylalanine	Phe	F	Hydrophobic
Glycine	Gly	G	Polar
Histidine	His	H	Charged/polar/hydrophobic
Isoleucine	Ile	I	Hydrophobic
Lysine	Lys	K	Charged (positive)
Leucine	Leu	L	Hydrophobic
Methionine	Met	M	Hydrophobic
Aspargine	Asn	N	Polar
Proline	Pro	P	Hydrophobic/polar
Glutamine	Gln	Q	Polar
Arginine	Arg	R	Charged (positive)
Serine	Ser	S	Polar
Threonine	Thr	T	Polar
Valine	Val	V	Hydrophobic
Tryptophan	Trp	W	Hydrophobic / partially polar
Tyrosine	Tyr	Y	Hydrophobic / partially polar

The translation of DNA basepairs to amino acids is of three to one. Since there are only four basepairs to code twenty amino acids, it is necessary to use more than one basepair in the coding. It is therefore necessary to use more than basepair to code an amino acid. Two basepairs are insufficient (there are $4^2 = 16$ possibilities) and three basepairs is "overdoing" it by suggesting the $4^3 = 64$ combinations. The number of combinations for three basepairs implies significant redundancy in coding the twenty amino acids and a STOP codon (the end of the protein chain). An argument in favor of degenerate coding is that it helps tolerate errors. The reading is indeed less sensitive to the substitution/mutations of basepairs. Nevertheless, the reading after deletions and insertions of basepairs may yield very different products.

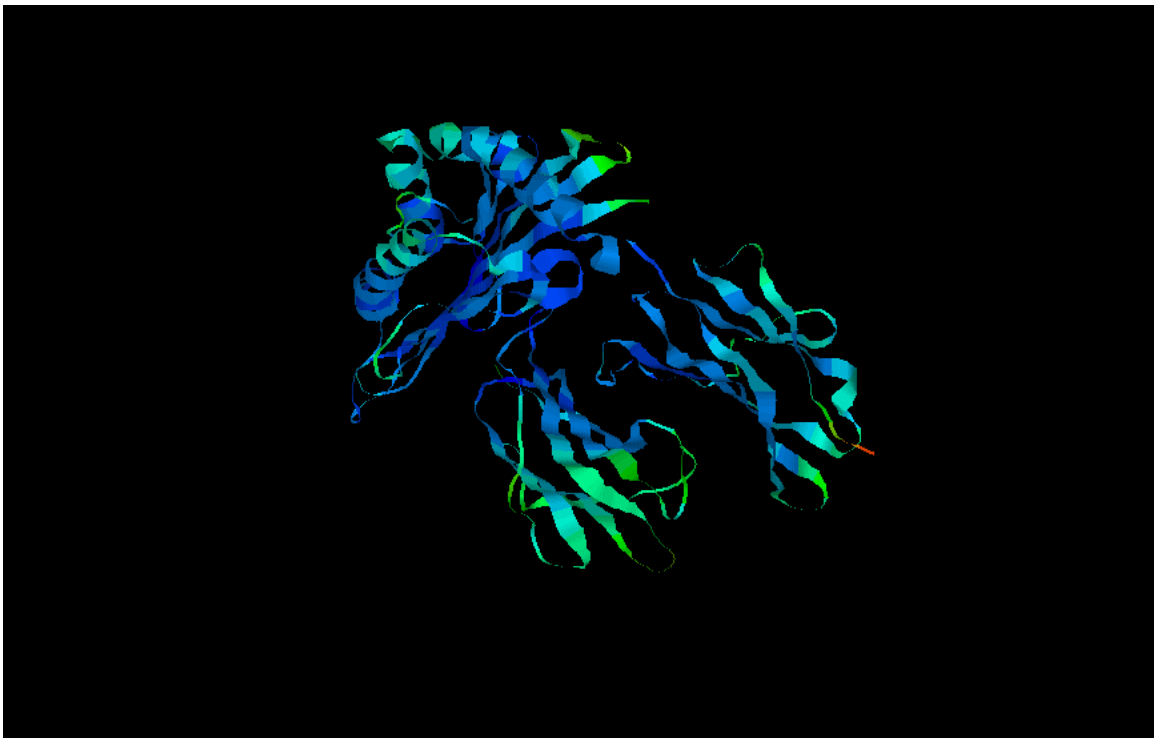
Our focus here is on the proteins. Throughout this text a sequence of a protein will be denoted by an italic capital letter, e.g., A. More explicitly, for a protein made of n amino acids, we write $A = a_1, ..., a_n$, where a_i denotes the amino acid at position i.

Proteins (typically) have compact shapes that support significantly larger number of structural motives (Figure 1) compared to DNA molecules.

MHC Class I Molecule B*5301 Complexed With Peptide Typdinqml From Gag Protein Of Hiv2
Smith, K. J., Reid, S. W., Harlos, K., McMichael, A. J., Stuart, D. I., Bell, J. I., Jones, E. Y.: Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity* 4 *pp.* 215 (1996)

Lengths of protein chains vary from a few tens to about a thousand amino acids. Very roughly we can divide the amino acids into two groups: hydrophobic (H) and polar (P), where we include the charged residues as a part of the polar group (table 2). This separation is useful in order to formulate conceptual models of proteins stability. Hydrophobic amino acids hate water and form a separate core with polar residues providing protective layer between water and the hydrophobic amino acids.

One realization of the protein-folding problem (the determination of structure from sequence) is problem of packing of heterogeneous one-dimensional chains (different amino acids at different spatial position) to maximize the contacts between hydrophobic residues and create a layer of polar residues between the water molecules and the hydrophobic core.



The core of the protein tends to be more hydrophobic

With the above simple picture in mind it is surprising that proteins cover a wide range of chain lengths. There are proteins of length of tens of amino acids and the longest single protein chains are in the thousands. With a few tens of amino acids it is difficult to construct hydrophobic core not exposed to water, and the thousands of amino acids may be an overkill to maintain just stability. Synthesis of thousand amino acids is more likely to include errors. Disposing an error of one thousand amino acid is more wasteful than disposing one hundred amino acids. Of course, proteins are made for more than stability, they suppose to perform many actions in the cell, (such as receiving and submitting signal, transporting material, and executing enzymatic reactions). More than a requirement to structural stability determines their sequences and shapes.

In a slightly different pose we may ask: According to what a protein is selected to perform a function in the cell? There are numerous determinants of protein sequences and

shapes, not the least important is pure chance. For example, stability, biological activity, and adaptation to different environments suggest a wide range of restrictions on protein design, native and non-native alike. In many cases it is necessary to investigate the problem including its environment and the sequence of an individual molecule will not provide sufficient information on its interactions within the cell, even if we will understand it to the quark level. The wide range of effectors makes the prediction of protein function from first principles (based on physical and chemical analysis of the single molecule) exceptionally difficult.

Consider the question, which is a starting point to many investigations in CMB, How can we determine protein function from the sequence of the amino acids S? A large fraction of the book describes approaches to address exactly that problem. One approach of studying function is to do it directly by modeling the physics and chemistry of their operation. Another approach is to seek evolutionary links to other related proteins. The first approach has the advantage of being less dependent on the availability of information on function of other proteins. However, it is significantly more expensive computationally and in has a lower success rate when the alternative approach works. The second method, relying on evolutionary links, is pretty good if an evolutionary related protein is found in databases of well-characterized proteins. In addition to the study of protein function the second approach is a tool to research evolutionary relationships between species. If no such link is found we are left with only the first method in our disposal. In that sense the first approach is more general.

In the present book we discuss both of these approaches, and in the next section we examine the information approach to function determination. We assume a bank of

protein sequences on which significant information is available and we wish to determine if the sequence of an unknown protein is related or similar to one of them.