1. Proteins Are Informational and Functional Biological Polymers

a. Proteins perform many of life's functions

Proteins are biological molecules that perform many of life's functions. Some proteins ***catalyze*** biochemical reactions, for example, oxidation of glucose to produce energy, copying of DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), repair of DNA, synthesis of proteins and other biomolecules, and degradation of toxins. Proteins that catalyze reactions are called *enzymes*. Some can have the opposite effect of blocking reactions; these are called ***inhibitors***. Some proteins are ***hormones***, and also ***receptors*** for these hormones, as well as other ***regulators*** of biological function. Some proteins act as switches for ***signal transduction***. Some proteins ***bind*** and ***transport*** biologically important substances such as ions, oxygen, glucose, lipids, and many other molecules, often from one cellular environment to another. Some proteins ***convert**, **transport** and* ***store energy***, for example in photosynthesis or in converting light to chemical signals in the eye. Some proteins **form *channels*** *and* ***gates***, embedded in membranes through which small molecules and ions enter and leave the cells. Many other proteins are essential elements of the ***immune*** system, such as the *immunoglobulins*.

b. Proteins are informational macromolecules

Each DNA molecule in the nucleus of each of the trillions of cells in our body can be viewed as a *database of information* because it carries the complete set of *instructions* for specifying a cell's structure, including the proteins utilized by the particular cell. *Triplets* of bases, termed *codons*, code for specific amino acids. Specific triplets serve for chain initiation (AUG), and others for termination (UAG, UAA and UGA). There are $4^3 = 64$ different possible base triplets, given four types of bases. This is in excess of the total number of amino acids. Most amino acids are thus specified by multiple codons, i.e., the genetic code is *degenerate*.

Table I.1.1 Amino acids naturally occurring in proteins

Amino acid Type	Three-letter code	Single letter abbreviation	Occurrence[(a)] (%)
Alanine	Ala	A	8.3
Alanine	Arg	R	5.7
Asparagine	Asp	N	4.4
Aspartic acid	Asp	D	5.3
Cysteine	Cys	C	1.7
Glutamine	Gln	Q	4.0
Glutamic acid	Glu	E	6.2
Glycine	Gly	G	7.2
Histidine	His	H	2.2
Isoleucine	Ile	I	5.2
Leucine	Leu	L	9.0
Lysine	Lys	K	5.7
Methionine	Met	M	2.4
Phenyl alanine	Phe	F	3.9
Proline	Pro	P	5.1
Serine	Ser	S	6.9
Threonine	Thr	T	5.8
Tryptophan	Trp	W	1.3
Tyrosine	Tyr	Y	3.2
Valine	Val	V	6.6

Question:

Do observed residue frequencies conform with random arrangements/mutations of nucleotides?'Hint: Use the values $p_0(U) = 0.220$, $p_0(A) = 0.303$, $p_0(C) = 0.217$ and $p_0(G) = 0.261$, and calculate the random probabilities for each amino acid. Compare with the results listed in Table 1.

In Figure I.1.1 the *observed* frequencies of amino acids (last column of *Table I.1.1*) are plotted against what is *expected* from random permutations of nucleotides. The expected or *random* frequencies are calculated as follows: Let us consider tyrosine, for example. Two codons, UAU and UAC, code for tyrosine (*Table I.1.2*). The random probability of the first is $p_0(UAU) = p_0(U)\, p_0(A)\, p_0(U)$, and that of the second is $p_0(UAC) = p_0(U)\, p_0(A)\, p_0(C)$, where $p_0(X)$ is the natural frequency of occurrence of nucleotide X in messenger RNA. Using the values $p_0(U) = 0.220$, $p_0(A) = 0.303$, $p_0(C) = 0.217$ and $p_0(G) = 0.261$, the total probability of the triplets UAU and UAC is calculated as 0.029. However, not all 64 codons specify amino acids: three of them are stop signals. After dividing by the approximate correction factor 61/64, the total probability of the codons for Tyr becomes 0.031. The percentage of Tyr residues expected from the composition of bases in RNA is therefore 3.1 %. This number is compared in *Figure I.1.1* with the actual fraction (3.2 %) observed {McCaldon & Argos 1988 ID: 289} in 1021 unrelated proteins.

The fit between the observed and expected frequencies is good (correlation coefficient 0.89 if Arg is disregarded). This suggests that a significant proportion of amino acids has arisen by *random* arrangements of DNA nucleotides. A large number of evolutionary changes in proteins may likewise have resulted from random changes in nucleotide

sequence,, which have no effect upon the fitness of the organism. Such genetic changes are referred to as *neutral mutations*. *Darwinian natural selection* for adaptive purposes probably does not screen out such functionally insignificant changes. It rather operates on the changes occurring at sites – occupied by evolutionarily *conserved* or *invariant* residues – which are critically important for function and/or stability.

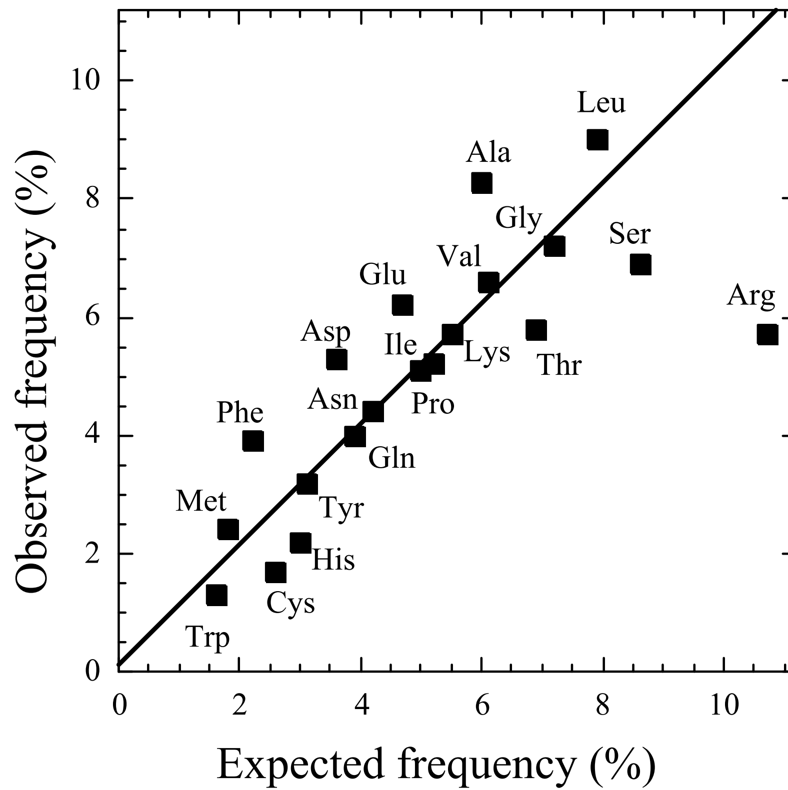

Figure I.1.1 *Similarity between the observed frequencies of amino acids and the frequencies predicted by the genetic code and random permutations of DNA nucleotides. The frequencies are given as percentages of amino acids. The best fit line from linear regression is shown.*

2. Genomes: From Sequence to Structure

a. A fundamental paradigm of protein science: amino acid sequence encodes the three-dimensional structure, which in turn specifies the function

The mapping between sequence and its native structure is often referred to as the *folding code*, or the *second genetic code*, in analogy to the (first) genetic code (*Table I.1.2*). As described above, the genetic code is a dictionary for translating the four-letter alphabet of nucleotides in DNA and RNA into the twenty-letter code of proteins. The folding code is different. The folding code is not a simple conversion of one-dimensional information in one alphabet into one-dimensional information in another alphabet, like the genetic code. Rather, the folding code is an expression of physical forces that cause a given one-dimensional sequence of amino acids to coil up or collapse into a particular three-dimensional structure.

According to the fundamental paradigm of protein science, the amino acid sequence specifies the three-dimensional structure, which in turn determines how the protein functions: **SEQUENCE —> STRUCTURE —> FUNCTION**. Thus, understanding how the three-dimensional structure is encoded in the amino acid sequence is an important step towards understanding the encoding of biological function in biomolecules and comprehension of complex organisms from their genome information.

b. There has been recently an exponential growth in sequence, structure and expression levels information from experiments

In the last few decades, there has been an explosion in the number of experimentally determined sequences and structures of biomolecules. Among these, sequence characterization of genomes, or *genome sequencing*, has been the fastest, while structure determination, impeded by the details of crystallization and data collection and analysis, has been slower. Compare the number of sequences accumulated in databases over the last 25 years, with the number of the three-dimensional structures elucidated at the same time, in *Figure I.2.2. {Rost 1998 ID: 331}*.

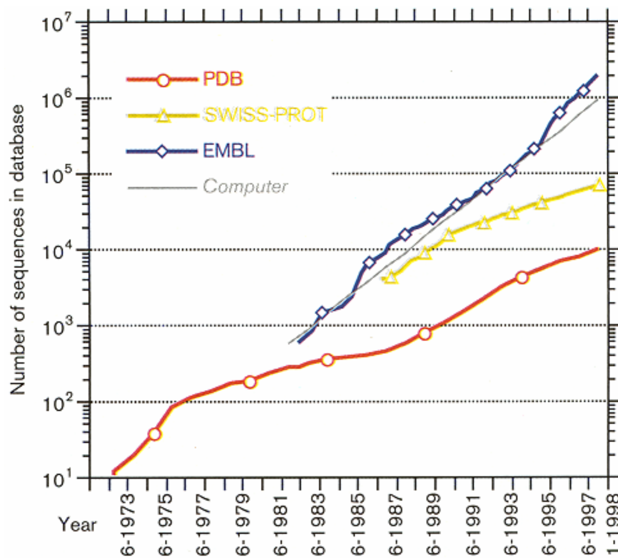

Figure I.2.2. *Increase in the numbers of known sequences a structures. PDB, the database of 3-dimensional protein structures is shown by the red curve. SWISS-PROT and EMBL are the protein and nucleotide sequence databases, respectively. The former is about one order of magnitude larger than the PDB, and the latter more than two orders of magnitude. The computational speed is shown by the thin gray line, which has grown at approximately the same pace the increase in the number of known sequences. (Figure 2 {Rost 1998 ID: 331}*)

The Protein Data Bank (PDB) is the database of known three-dimensional structures of proteins and their complexes {Abola, Bernstein, et al. 1987 ID: 26}{Bernstein, Koetzle, et al. 1977 ID: 332} {Berman, Westbrook, et al. 2000 5 /id}. In the PDB, atomic coordinates down to nanometer resolution are available for > 17,500 structures. The nucleotide sequence database (EMBL) {Stoesser, Sterk, et al. 1997 ID: 334} is about one order of magnitude larger than the amino acid sequence database (SWISS-PROT) {Bairoch & Apweiler 1998 ID: 333}. In fact, knowledge of nucleotide sequence does not necessarily provide information on the amino acid sequence, in spite of the well-established genetic code (*Table I.1.2*). The discrepancy is due to the fact that the *protein-encoding sequences*, comprise only a few per cent of the genome and an average of about 5% of each gene {Lander, Linton, et al. 2001 8 /id}. Only 1.1% of the human genome is spanned by exons, 24% by introns, the remaining 75% being intergenic {Venter, Adams, et al. 2001 1181 /id}.

The output from X-ray experiments is now > 2000 new structures per year (Figure I.2.3), although the fraction of 'new folds' is decreasing.

The number of experimentally determined structures continues to lag behind that of known sequences, given the much faster methods of gene sequencing and the experimental difficulties in structure determination. In view of this gap, **prediction of unknown structures is one of the major goals of computational biologists.** This goal is now more tractable with the rapid increase in computing power, and the access to the rapidly growing ensemble of structures that may serve as templates for assigning structures, or may provide insights for constructing realistic models.

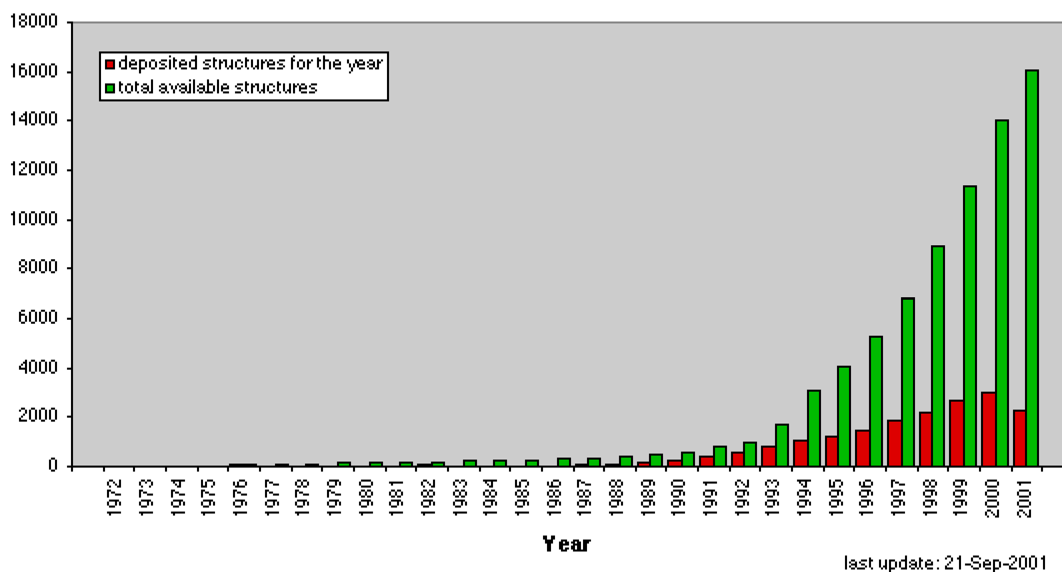

Figure I.2.3. *Increase in the size of the PDB (from http://www.rcsb.org/pdb/holdings.html#holdings)*

c. Experiments can also probe folding mechanisms and functional motions, and foster the construction of more accurate models for computer-aided studies

In addition to *equilibrium* structure determination, we now have direct access to the *changes* in structures, or structural *dynamics* or *mechanics*, controlling several processes.

Single molecule observation and manipulation have now come of age {Mehta, Rief, et al. 1999 ID: 319}. The use of fluorophores or other site-specific labels, together with rapid data acquisition methods that maximize the signal-to-noise ratio, now minimize the possible masking effect of ensemble averages {Weiss 1999 ID: 320}. *Figure I.2.4* illustrates the use of fluorescence spectroscopy

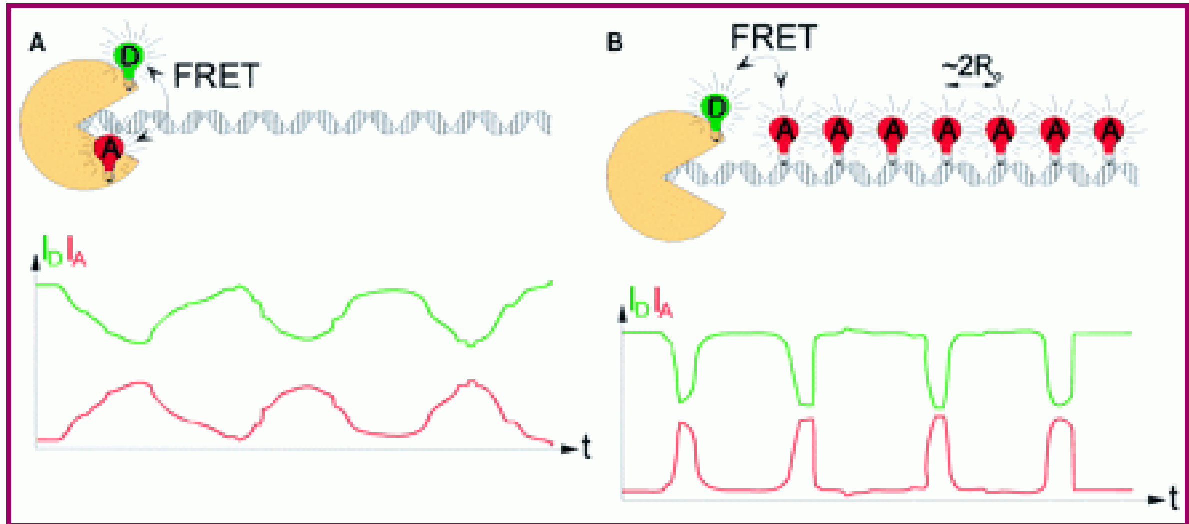

Figure I.2.4. Cartoon illustrating the use of fluorescence energy transfer (FRET) for monitoring the time evolution of (A) intramolecular motions of a nuclease, and (B) the diffusion of the nuclease along the DNA molecule. In part (A) the donor and acceptor fluorophores are both affixed on the enzyme. Their separation changes with the conformational state of the enzyme during its catalytic activity. The emissions from the two fluorophores shown by the lower curves reflect the period of the catalytic cycle. There is a nonradiative energy transfer between the acceptor and the donor whenever these groups are close enough, resulting in a periodic decrease in the observed (radiative) fluorescent emission as a function of time. In part B, the donor and acceptors are appended to the protein and DNA molecule, respectively. Likewise, there is a signal whenever the chromophores come into close contact (Förster radius refers to the distance at which 50% of energy is transferred). The accompanying emission spectra reflect the diffusion rate of the enzyme along the(DNA molecule (from {Weiss 1999 ID: 320}, Figure5).

d. A major challenge to the structural community: Structural Genomics

The Human Genome (HG) Project started as an international collaborative project aiming at completely *sequencing* or *mapping* the human genome, i.e. developing maps that depict the order in which genes and other DNA landmarks are organized along the DNA molecules, with the major objective of identifying genes associated with diseases. The first draft of the HG has been recently completed {Venter, Adams, et al. 2001 1181 /id}{Jasny & Kennedy 2001 1 /id}.

Presently the genomes of 44 organisms have been completely sequenced (Dec 2001). Information on the genome of about 800 organisms can be found in Entrez Genomes http://www.ncbi.nlm.nih.gov/Genomes/index.html. These represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses and organelles. Transmembrane proteins - which are usually non-globular, non-crystallizable and insoluble in aqueous environment - represent as much as 25-30% of a genome. These and other unknown or hypothetical genes amount to up to 70% of the mapped genomes, with the remainder being structurally or functionally known proteins. It is a major

challenge to the structural community to determine what these unknown structures are.

Interestingly, for the ~16,700 PDB structures currently available, the number of distinct *folds* (3-dimensional structures at a coarse-grained scale, also referred to as *architecture*) amounts to less than 700. In fact, sequence tends to diverge more rapidly than structure (see *Figure I.2.5*). Many proteins encoded by the newly sequenced genes are indeed found to fold into known architectures, when their 3-D structures are subsequently determined.

On the other hand, a distinct fold implies a different mechanism of action, or functional motion. And, in order to acquire maximal new information about structure and function, one should carefully select the genes and/or organisms to be investigated. Computations can and should provide guidance to experimentalists about which sequences are most likely to yield 'new' folds. Another feature worth noting is that the number of distinct folds is small enough so that *all* known folds can be thoroughly examined by computations for inferring mechanisms of functional motions imprinted in each fold.

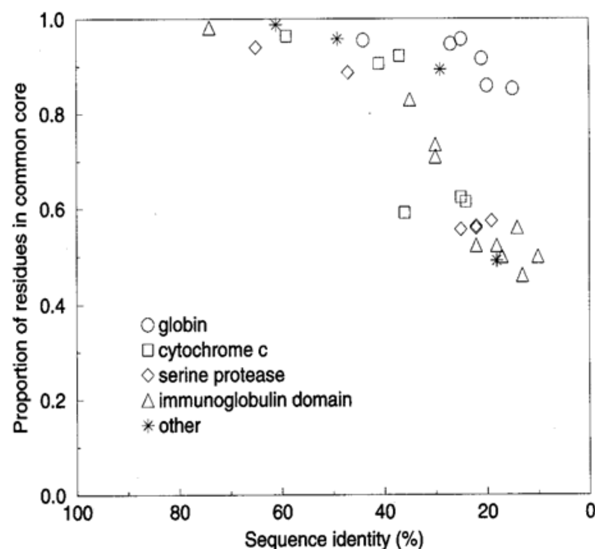

Figure I.2.5. *Fraction of residues superimposable in the core of the structures as a function of sequence identity percentage. The majority of core residues are structurally superimposable, for example, for two proteins sharing > 30 % sequence identity. The figure is redrawn ({Durbin, Eddy, et al. 1998 ID: 305}, Figure 6.2) from data reported by Chothia and Lesk {Chothia & Lesk 1986 ID: 310}. 'Other' refers to two dehydrofolate reductases, two lysozymes, and the pairs plastocyanin/azurin and papain/actinidin*

Even though significant efforts have been focused on gene sequencing and computer-aided sequence analysis within the scope of the Human Genome Project, an ultimate goal of structural biology and bioinformatics, and perhaps that of the biomedical and life sciences, is to understand, control and to be able to design biological *function*. Because of the close connection between a protein's structure and the way it operates, the analysis of *structures* is a more powerful way of recognizing function than *sequence* analysis. In recognition of this fact, a coordinated, international endeavor, in ***Structural Genomics*** has been initiated {Gaasterland 1998 ID: 326}{Kim 1998 ID: 325}{Pennisi 1998 ID: 324}{Shapiro & Lima 1998 ID: 328}{Brenner, Barken, et al. 1999 ID: 327}, along with *Functional Genomics*. Within the scope of these enterprises, efforts should focus on one or more selected organism whose genome has been completely mapped, and the structures of *all* proteins encoded by the investigated genome would be found by experimental and computational means.

In addition, there is some possibility that the organization of genes may aid us in learning about protein functions, i.e. to estimate function from the sequence of complete genomes, in the absence of structural knowledge. The organization of related sets of proteins into operons - genetic units formed by contiguous genes that encode proteins involved in a given function - is well known. For example, DNA sequences coding for the proteins involved in the synthesis of lactose are positioned close together in the so-called *lac* operon. It has been further postulated that proteins that interact are likewise closely placed in the genes, and such cases can be detected by locating the recombination of domains of two separate proteins into a single protein existing within other organisms {Marcotte, Pellegrini, et al. 1999 ID: 273}. Other experimental means, such as the two hybrid method, have been developed to determine directly which proteins interact with other proteins, and extensive catalogs of these functioning networks of interacting proteins have been developed.

e. Distribution of protein categories from genomes

Recently, the 2.91-billion base pair consensus sequence of the first draft of the human genome (HG) was released {Venter, Adams, et al. 2001 1181 /id} {Lander, Linton, et al. 2001 8 /id}. The International Human Genome Sequencing Consortium found evidence for 29,691 human transcripts {Lander, Linton, et al. 2001 8 /id}, while the commercial project of Celera Genomics founded 39,114 genes {Venter, Adams, et al. 2001 1181 /id}. A more recent consolidation of HG transcripts and protein database by Yuan and coworkers suggests that humans may have more than 70,000 genes (F.A. Wright et al., "A draft annotation and overview of the human genome" *Genome Biology*, 2:1-18, 2001).

Figure 1.2.7 gives an overview of the distribution of HG proteins based on their function. Figure 1.2.8 shows the number of proteins in distinct functional categories for five essentially completed eukaryotic proteomes (human, yeast, worm, mustard weed and fly).

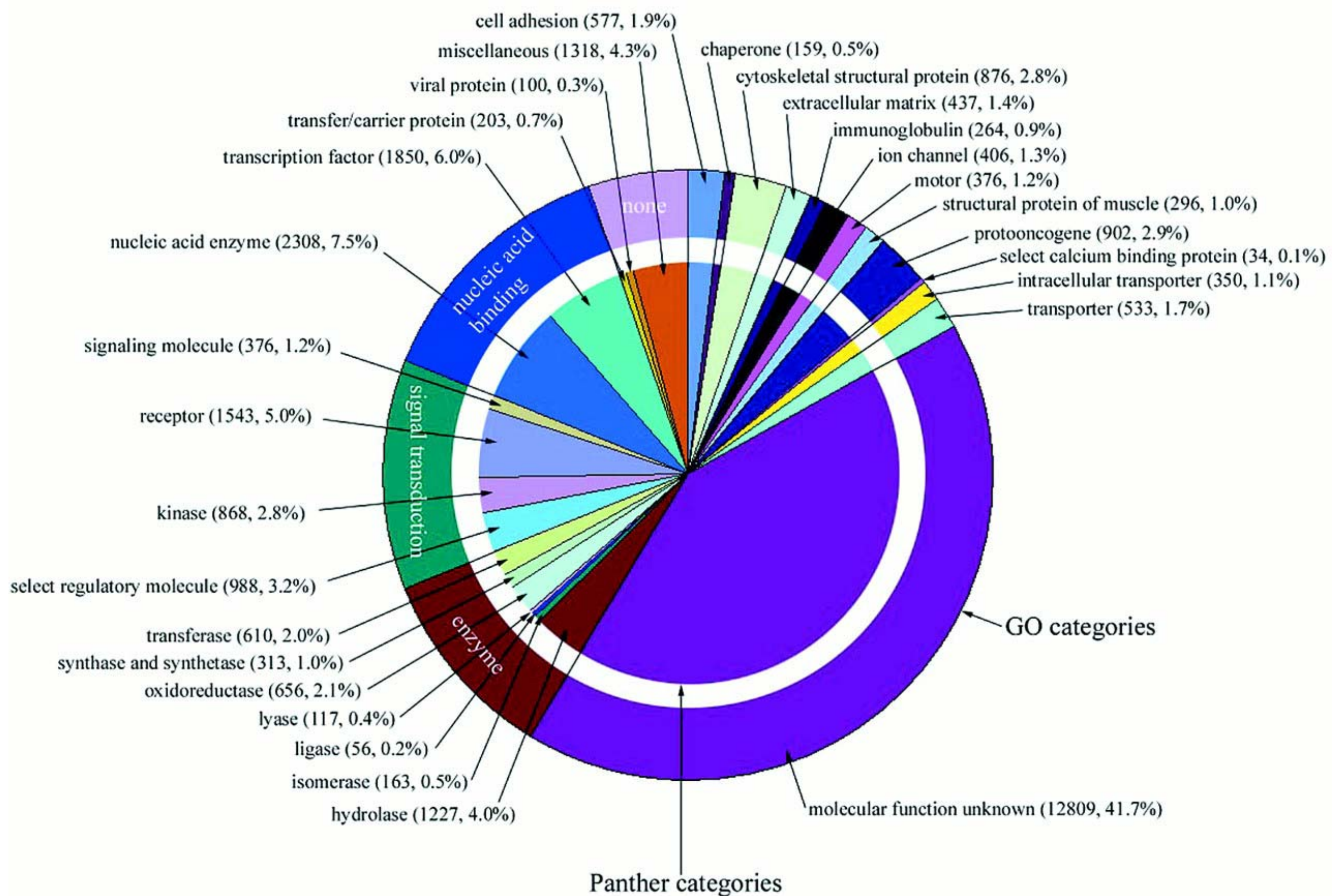

***Figure 1.2.7**. Distribution of the molecular functions of the 26,383 human proteins. Each slice lists the numbers and percentages (in parentheses) of gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. **25:** 25-29), and the inner circle shows the assignment to Celera's Panther molecular function categories. (Figure 15 in {Venter, Adams, et al. 2001 1181 /id})*

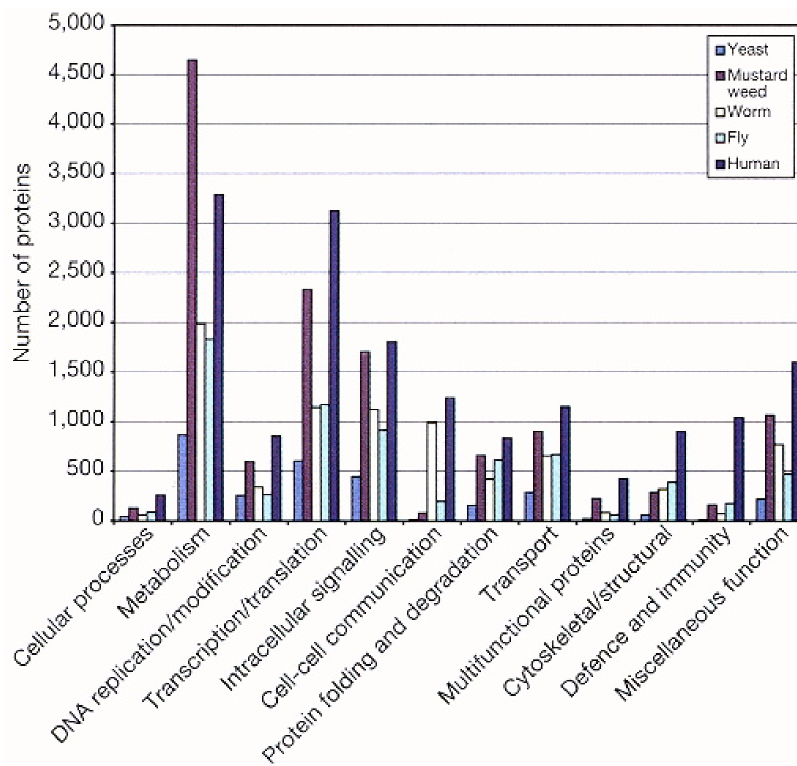

Figure 1.2.8. Functional categories in eukaryotic proteomes. The categories were derived from functional classification systems, including the Gene Ontology project. (Figure 37 in {Lander, Linton, et al. 2001 8 /id}

3. Proteins are Microscopic Machines

a. Proteins have large numbers of conformational degrees of freedom

Proteins are macromolecules. The most fundamental property distinguishing macromolecules from small molecules is presumably their ability to assume a large number of *internal configurations*. This feature implies that statistical approaches, or statistical mechanical and thermodynamic theories, are needed for understanding the behavior of macromolecules at the molecular level. The *macroscopic* properties of macromolecules indeed represent the average behavior of large numbers of *conformational* states.

b. Yet, each protein folds into a *unique* biologically active configuration, its *native* state, under physiological conditions

Proteins are special because, despite the enormous number of available configurations, under physiological conditions each globular protein strikingly assumes a *unique* three-dimensional or *tertiary* structure, its *native* (N) or *folded* (F) configuration. The physiological environment refers to a rather concentrated aqueous system, with numerous components present such as metal ions or prosthetic groups, at least.

c. Many, if not most of our diseases have their origins in our genes

Genes may be altered, or *mutated* in many ways, whether in the essential inheritable nuclear DNA (*germline mutation*) or only in a specific cell or set of cells (*somatic mutation*). The most common type of mutation is a base *mismatch*, - a misspelling. A base replaces another in this case. Others are *deletions* and *insertions*. The mutant protein can function perfectly or imperfectly, or can be totally disabled. This response may be critical to an organism's well being if this protein plays a vital role.

Gene mutations can be either *inherited*, or *acquired*. Hereditary mutations are carried in the DNA of reproductive cells. Acquired mutations, on the other hand, develop during the individual organism’s lifetime. Such mutations often arise under the influence of extreme environmental effects, such as radiation, chemicals or other harsh conditions. Actually, mistakes occur quite frequently in the DNA. But, there exist several remarkable mechanisms to *repair* genes, especially during cell division. If these mechanisms fail to operate properly, then mutations are passed along in subsequent copies. DNA repair and damage response mechanisms play a vital role in maintaining the integrity of the genome against thousands of mutagenic and cytotoxic modifications daily induced in the DNA sequence {Wood, Mitchell, et al. 2001 17 /id}{Lindahl & Wood 1999 22 /id}.

Most genomic variation is attributable to single nucleotide polymorphisms (SNPs). More than 1.42 million SNPs have been assembled during the course of the HG project, yielding an average density of one SNP every 100-300 base. Among these, 60,000 fall within exons (encoding and untranslated regions) {Sachidanandam, Weissman, et al. 2001 1182 /id}. Human diseases known to stem from altered genes are compiled in the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/Omim/). Common disorders such as heart diseases and most cancers result from a complex interplay between multiple gene defects and environmental factors. Cancerous cells are otherwise normal cells that proliferate without restraints when the genes that control their

orderly *replication* and *growth* are mutated, - or the proteins encoded by these genes are misfolded.

d. The range of effects of mutations, and their functional and medical implications, are extremely broad

The ability of proteins to adapt to structural perturbations is termed their *structural plasticity*. Yet, some selected mutations can induce large conformational changes, or replace critical residues in the active sites that are required for enzyme function, and cause unexpected effects. Among these can be changes to the interaction of the protein with other proteins, caused either by destroying requisite intermolecular binding sites or by introducing new deleterious binding sites. The effect of mutations indeed depends strongly on the sequence and environment, a factor referred to as *context*.

Figure I.3.2 illustrates, for example, the change in conformation induced in *Arc repressor* homodimer - inhibitor of enzymatic transcription, upon interchanging the sequence positions of two of its residues, Leu12 and Asn11 {Cordes, Walsh, et al. 1999 ID: 293}. The mutant protein exhibits a significant structural change, - from a double stranded β-sheet to a pair of right-handed helices. Compare the wild type and mutant structures shown in parts *A* and *C*. Parts *B* and *D* present an alternative view of the two structures. The parts of the dimer that undergo structural changes are shown in blue, and the remainder in red.

The significant conformational change induced by interchanging the sequential position of two residues in *Arc repressor* can be explained as follows: Leu is hydrophobic (H), and Asn is polar (P). The interchange of these two residues changes the pattern of the original sequence Gln9-Phe10-Asn11-Leu12-Arg13-Trp14 in each monomer from PHPHPH in the wild type protein to PHHPPH in the mutant. The former is typical of β-strand sequences, in which alternating polar residues are exposed to solvent; whereas the latter reflects the structural periodicity of a helix with three to four residues per turn. Such simple changes to the binary pattern of H and P residues provide one mechanism for the evolution of protein folds.

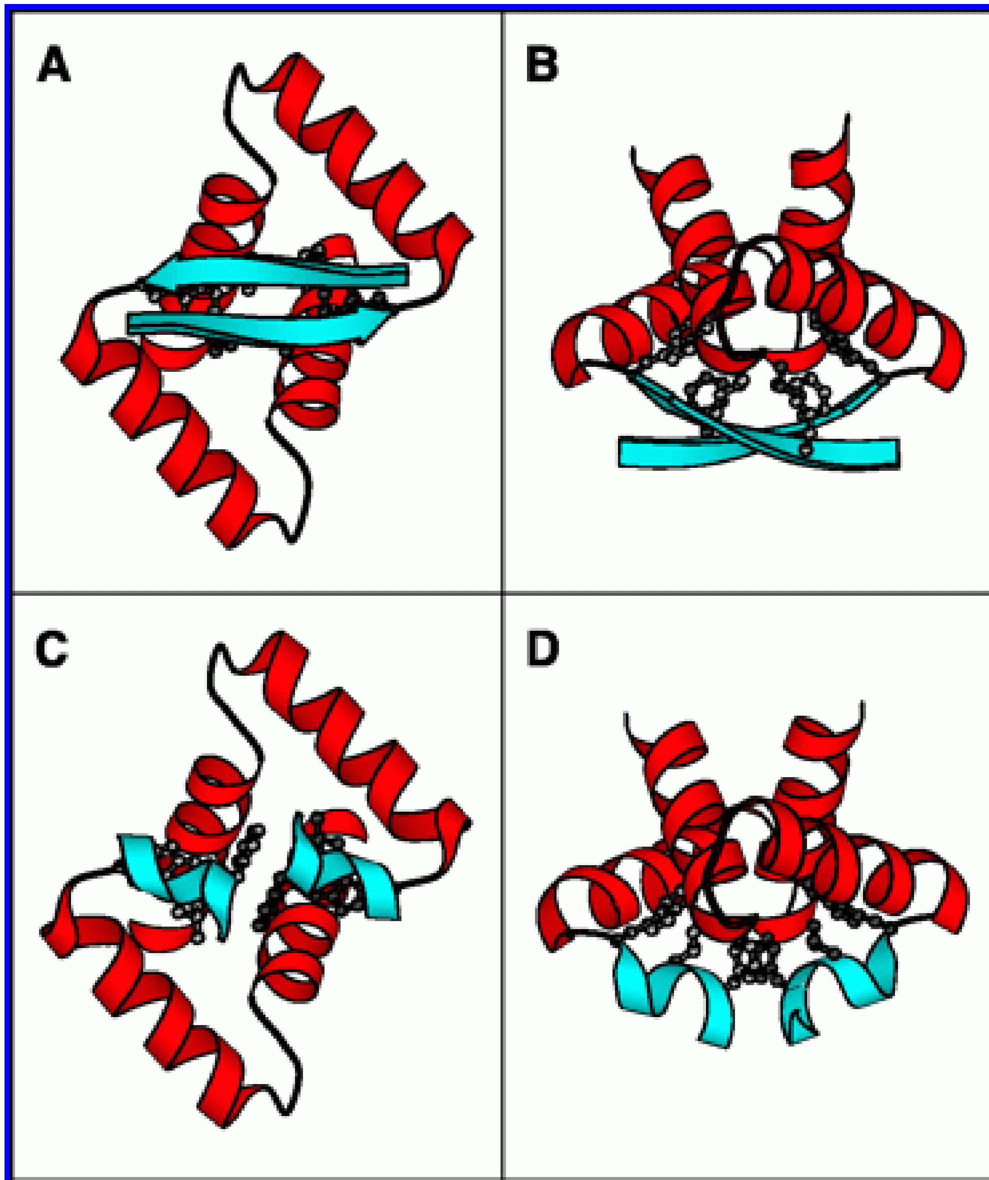

Figure I.3.2. *Structural change in the β-sheet region of the Arc repressor homodimer induced upon interchanging two adjacent residues, Asn11 and Leu12, in each monomer. Parts A and B display the structure of the wild type protein from two different views, and C and D are the corresponding views of the mutant protein structure. The two β-strands contributed by one monomer, each, in the wild type protein are repacked as helices in the mutant to allow the protein to maintain solvent exposure of the polar residue Asn11 and burial of hydrophobic residue Leu12 (Fig 4 of {Cordes, Walsh, et al. 1999 ID: 293}).*

e. Protein misfolding and aggregation cause diseases

A radical change in conformation usually entails a loss of activity. Such "switch" mutants may be important in causing certain diseases. For example, inherited disorders such as familial *Creutzfeldt-Jacob's* disease (CJD) or fatal familial insomnia are putatively linked to mutations in a small protein, *prion pr*otein (PrP) {Prusiner 1994 ID: 297}{Prusiner 1997 ID: 301} (*Figure I.3.3*). The posttranslational modification and consequent accumulation of PrP is also responsible for animal diseases such *as bovine spongiform encephalopathy* (BSE, also known as mad cow disease) and sheep *scrapie* (Sc). The normal cellular protein PrP^C is proposed {Pan, Baldwin, et al. 1993 ID: 299}{Huang, Gabriel, et al. 1994 ID: 298} to switch in these cases from a predominantly helical structure to a modified isoform (PrP^{Sc}) with a high β-sheet content (*Figure I.3.3)*. The latter is relatively insoluble and resistant to digestion by proteases. It forms pathogenic *aggregates*, or *amyloid fibrils*, responsible for central nervous system degenerative disorder.

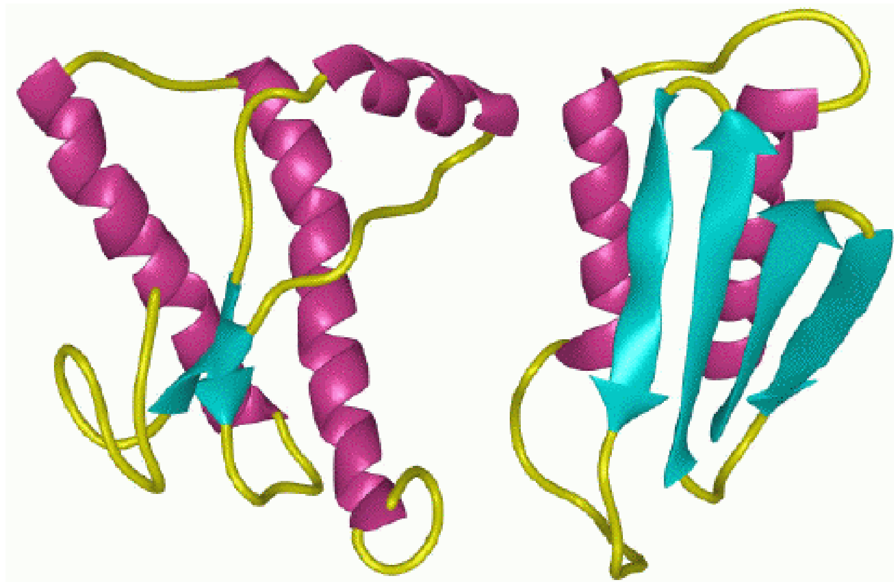

Figure I.3.3. *NMR structure of rPrP(90-231), residues 90-231 of the recombinant prion protein rPrP from Syrian hamster, on the left, compared to the structure of the disease-causing form,* PrP^{Sc} *proposed by Huang, Prusiner and Cohen {Huang, Prusiner, et al. 1996 ID: 300}, on the right. The NMR structure closely resembles that of the normal cellular protein* PrP^C*. Major structural change occurs at the N-terminal portion, where the relatively disordered structure is organized into a four-stranded antiparallel sheet. (from http://www.cmpharm.ucsf.edu/home/~wallacea/public_html/prion/nobel.html)*

f. An important step in controlling biological activities is to understand signaling and control mechanisms

Activation can be triggered in several ways. One common way is binding at one site that affects binding or activity at another distant site. This is termed *allostery*. In a number of complexes, binding of a substrate, a ligand or simply an ion, induces one or more conformational changes, which is communicated through cooperative effects to biologically active sites. Thus, biological action is triggered by a conformational transition, i.e. by a *physical* change. One example is the binding of Ca^{2+} to calmodulin (CAM), a protein that participates in numerous regulatory processes. CAM is a dumbbell-shaped protein comprising two domains, each of which contain two helix-loop-helix motifs that bind Ca^{2+} ions, called EF hands *(Figure 1.3.5).* The binding of Ca^{2+} induces a conformational change which exposes an otherwise buried methionine-rich hydrophobic patch; and the latter, in turn, binds with high affinity to many CAM-binding proteins whose activities are modulated by CAM. An example of a CAM-binding protein is myosin light chain kinase, an enzyme that phosphorylates and thereby activates the muscle protein myosin.

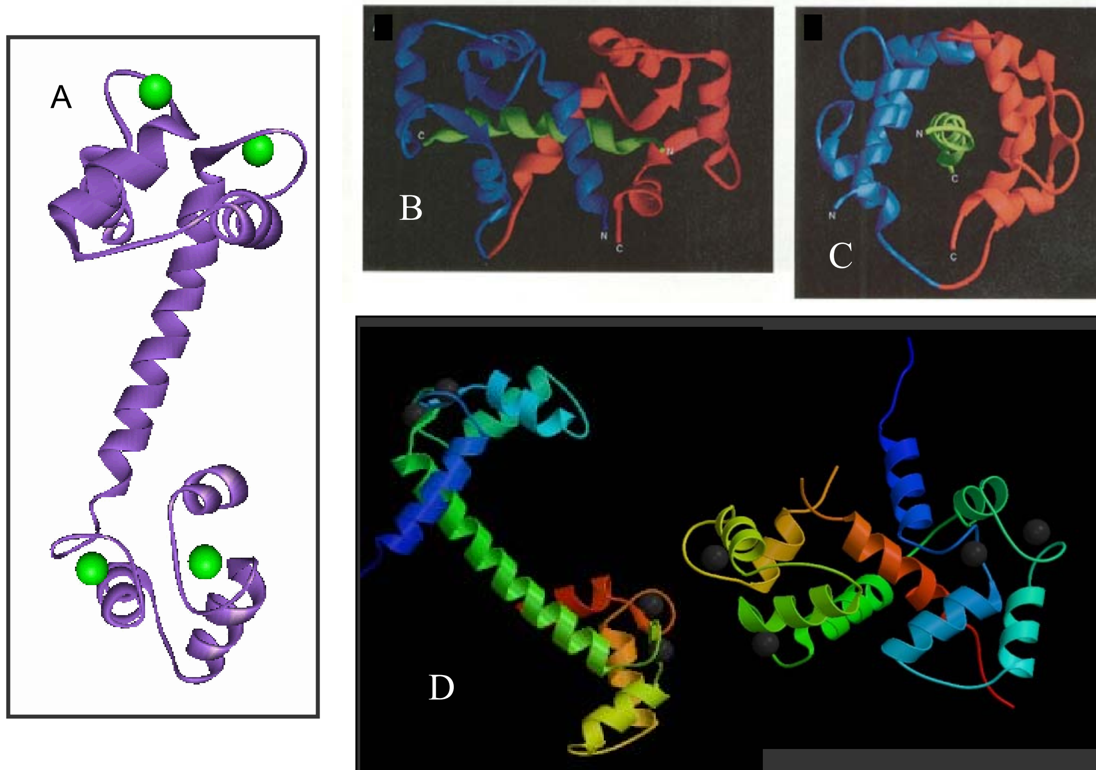

Figure I.3.5. *(A) Structure of calmodulin (CAM). CAM has two lobes connected by a 21 residue central helix. Each lobe binds two calcium ions (green) (*http://www-structure.llnl.gov/Calmod/calmod.html*). (B-C). Side view (B) and top view (C) of CAM complexed with a peptide (myosin light chain kinase), determined {Ikura, Clore, et al. 1992 1183 /id} by NMR. The peptide is shown in green, the N and C-terminal halves of CAM are colored blue and red* (from *(D) Comparison of the two conformations of CAM deposited in the PDB by {Babu, Bugg, et al. 1988 1184 /id}(crystal structure, left) and {Ikura, Clore, et al. 1992 1183 /id}(NMR structure, right). Sequential residues' colors change from blue to red.*