BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJ.)
FIRST SEMESTER 2016-2017
BIO F417, BIOMOLECULAR MODELING
COMPREHENSIVE EXAMINATION
TOTAL WEITAGE 35% Date: 08.12.2017 DURATION: 3Hrs. (Part A & Part B)
Total Marks (47+23) =70

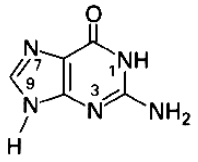
- **Answer Part A and Part B in separate answer sheets.**
- **Irrelevant answer may attract penalty.**

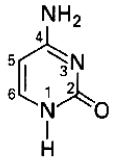
PART – A (CLOSED BOOK) (Max. duration: 2 Hrs., Max. Marks 47)

1. a) The following structural parameters are from three different forms of DNA double helix. Identify them with proper justification. In each form, values are given for CG/GC pairs and CpG/GpC steps. All symbols are representing standard DNA notation. **[3]**

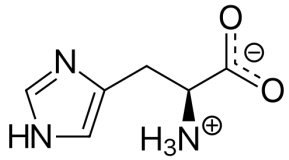
	β	δ	P	χ	Inclin	Rise	X-disp	Prop	Slide
Form-1	**-146°**	**156°**	**192°**	**-98°**	**-6°**	**3.38Å**	**-0.7Å**	**-12°**	**0.23 Å**
Form-2	**-176°**	**140°**	**156°**	**-161°**	**7°**	**3.7Å**	**-1.2Å**	**2.6°**	**0.5 Å**
Form-3	**-152°**	**82°**	**38°**	**-154°**	**19°**	**2.56Å**	**-5.4Å**	**-11°**	**-1.5 Å**

b) Draw all possible hydrogen bonded basepair with following two bases (consider inter as well as intra nucleotide basepair) where at least two hydrogen bonds exist between them. **[4]**


guanine


cytosine

2. a) From all possible rotameric conformations of given histidin (His) molecule (shown below), choose favorable combination of rotameric conformation. Show (by Newman projection) all possible rotameric states of His. **[4]**



b) Draw atomistic hydrogen bond scheme of a parallel and an antiparallel B-sheet. Each chain should contain at least four residues. **[4]**
c) Between protein and DNA, which molecule is hard to simulate and why? **[2]**

3. a) Explain how *ab-initio* modeling technique can assist comparative modeling and threading technique. **[2]**
b) In the scenario of downhill protein folding, what would be "speed limit" of the folding process? **[2]**

c) List down all the typical criteria of selecting protein template of comparative modeling technique. **[3]**
d) Mention the utilization of following in Biomolecular modeling: **[3]**
i) Energy minimization **ii)** Protein folding theories **iii)** CATH database.

4. a) The success of molecular mechanics and molecular dynamics techniques for biomolecular simulation is heavily dependent on certain assumptions. Mention all assumptions which are considered during simulation. Explain the need for each of these assumptions. **[3+2]**
b) Explain why molecular mechanical energies are relative in nature. **[2]**
c) Compare molecular dynamics (MD) and monte carlo (MC) simulation techniques. **[3]**
5. Write notes on: **i)** Forcefield and its importance **ii)** β-turns **iii)** Template free modeling **iv)** Proein-DNA interaction **v)** Importance of RNA structure **[5X2=10]**

**********Good Luck**********

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJ.)
FIRST SEMESTER 2017-2018
BIO F417, BIOMOLECULAR MODELING
COMPREHENSIVE EXAMINATION
TOTAL WEITAGE 35% Date: 08.12.2017 DURATION: 3Hrs. (Part A & Part B)
Total Marks (47+23) =70

- **Answer Part A and Part B in separate answer sheets.**
- **Irrelevant answer may attract penalty.**

PART – B (OPEN BOOK) (Max. duration: 1 Hr., Max. Marks 23)

1. a) What is the protein folding code? **[3]**
b) What is the protein folding mechanism? **[3]**
c) Can we predict the native structure of a protein from its amino acid sequence? **[4]**
[**Hints:** You can also take help of the attached text which is a part of a review article related to protein folding.]

2. a) The attached table is omega (ω), phi (ϕ) and psi (ψ) torsion angles of amino acid residues which are part of a stable protein structure. From the torsion angle values draw a schematic diagram of the structure of given polypeptide chain indicating the residue number on each secondary structure unit. Justify your schematic diagram. **[3]**
b) Among all standard basepair and step parameters of DNA double helix, identify four most closely linked parameters. (pairwise) **[4]**
c) Discuss the salient features of Nussinov's RNA secondary structure prediction algorithm. **[3]**
d) Suppose during protein structure modeling through comparative modeling technique, your top template hit had 25% query coverage with 35% sequence identity. What would be your approach to complete modeling assignment? [Discuss all possible solution]. **[3]**

***************************Good Luck***********************************

Table

Residue	ω in °	ϕ in °	ψ in °
ASN2	-173.99	-51.454	-65.440
SER3	-178.50	-62.577	-44.878
ALA4	-174.16	-59.891	-46.702
GLU5	178.945	-63.208	-38.764
ALA6	176.810	-67.222	-38.954
TRP7	172.112	-63.280	-41.203
TYR8	175.886	-59.091	-49.220
ASN9	178.579	-61.952	-33.530
LEU10	174.090	-63.991	-43.865
GLY11	176.416	-59.902	-42.815
ASN12	176.266	-57.778	-42.208
ALA13	-176.48	-64.891	-39.852
TYR14	178.723	-73.031	-37.995
TYR15	172.751	-55.019	-45.590
LYS16	-178.19	-68.693	-21.528
GLN17	168.306	-78.573	-9.148
GLY18	-175.63	76.944	23.135
ASP19	-178.02	-104.42	74.630
TYR20	-168.56	-69.066	-33.789
ASP21	178.630	-60.604	-47.615
GLU22	178.313	-65.006	-43.649
ALA23	179.395	-55.050	-45.742
ILE24	179.696	-59.107	-44.765
GLU25	178.837	-56.646	-44.873
TYR26	-174.78	-76.009	-38.751
TYR27	172.832	-62.901	-37.046
GLN28	176.908	-62.634	-42.001
LYS29	176.730	-63.377	-43.811
ALA30	176.001	-55.408	-44.230
LEU31	-173.20	-75.246	-27.412
GLU32	170.854	-61.469	-42.547
LEU33	-178.48	-79.284	-30.765
ASP34	179.039	-148.47	86.817
PRO35	-177.96	-72.358	-2.407
ASN36	178.943	-92.049	5.109
ASN37	-179.38	-82.068	90.102
ALA38	-178.22	-58.556	-39.109
GLU39	-177.24	-66.227	-29.525
ALA40	176.451	-68.514	-41.801
TRP41	174.690	-58.520	-44.418
TYR42	-177.74	-62.856	-46.494
ASN43	-179.36	-70.629	-32.826
LEU44	170.396	-58.870	-45.771
GLY45	174.505	-53.422	-48.092
ASN46	-179.65	-59.786	-41.158
ALA47	179.926	-61.265	-43.002
TYR48	176.238	-66.117	-41.184
TYR49	173.069	-52.244	-48.311
LYS50	179.937	-63.164	-29.193

Review article on Protein Folding

****

INTRODUCTION

The protein folding problem is the question of how a protein's amino acid sequence dictates its three-dimensional atomic structure. The notion of a folding "problem" first emerged around 1960, with the appearance of the first atomic-resolution protein structures. Some form of internal crystalline regularity was previously expected, and α-helices had been anticipated by Linus Pauling and colleagues, but the first protein structures of the globins had helices that were packed together in unexpected irregular ways. Since then, the protein folding problem has come to be regarded as three different problems: (a) the folding code: the thermodynamic question of what balance of inter atomic forces dictates the structure of the protein, for a given amino acid sequence; (b) protein structure prediction: the computational problem of how to predict a protein's native structure from its amino acid sequence; and (c) the folding process: the kinetics question of what routes or pathways some proteins use to fold so quickly. We focus here only on soluble proteins and not on fibrous or membrane proteins.

WHAT BALANCE OF FORCES ENCODES NATIVE STRUCTURES?

Anfinsen's Thermodynamic Hypothesis

A major milestone in protein science was the thermodynamic hypothesis of Christian Anfinsen and colleagues. From his now-famous experiments on ribonuclease, Anfinsen postulated that the native structure of a protein is the thermodynamically stable structure; it depends only on the amino acid sequence and on the conditions of solution, and not on the kinetic folding route. It became widely appreciated that the native structure does not depend on whether the protein was synthesized biologically on a ribosome or with the help of chaperone molecules, or if, instead, the protein was simply refolded as an isolated molecule in a test tube. There are rare exceptions, however, such as insulin, α-lytic protease, and the serpins, in which the biologically active form is kinetically trapped. Two powerful conclusions followed from Anfinsen's work. First, it enabled the large research enterprise of in vitro protein folding that has come to understand native structures by experiments inside test tubes rather than inside cells. Second, the Anfinsen principle implies a sort of division of labor: Evolution can act to change an amino acid sequence, but the folding equilibrium and kinetics of a given sequence are then matters of physical chemistry.

One Dominant Driving Force or Many Small Ones?

Prior to the mid-1980s, the protein folding code was seen a sum of many different small interactions such as hydrogen bonds, ion pairs, van der Waals attractions, and water-mediated hydrophobic interactions. A key idea was that the primary sequence encoded secondary structures, which then encoded tertiary structures. However, through statistical mechanical modeling, a different view emerged in the 1980s, namely, that there is a dominant component to the folding code, that it is the hydrophobic interaction, that the folding code is distributed both locally and nonlocally in the sequence, and that a protein's secondary structure is as much

a consequence of the tertiary structure as a cause of it. Because native proteins are only 5–10 kcal/mol more stable than their denatured states, it is clear that no type of intermolecular force can be neglected in folding and structure prediction. Although it remains challenging to separate in a clean and rigorous way some types of interactions from others, here are some of the main observations. Folding is not likely to be

dominated by electrostatic interactions among charged side chains because most proteins have relatively few charged residues; they are concentrated in high-dielectric regions on the protein surface. Protein stabilities tend to be independent of pH (near neutral) and salt concentration, and charge mutations typically lead to small effects on structure and stability. Hydrogen bonding interactions are important, because essentially all possible hydrogen-bonding interactions are generally satisfied in native structures. Hydrogen bonds among backbone amide and carbonyl groups are key components of all secondary structures, and studies of mutations in different solvents estimate their strengths to be around 1–4 kcal/mol or stronger. Similarly, tight packing in proteins implies that van der Waals interactions are important.

However, the question of the folding code is whether there is a dominant factor that explains why any two proteins, for example, lysozyme and ribonuclease, have different native structures. This code must be written in the side chains, not in the backbone hydrogen bonding, because it is through the side chains that one protein differs from another. There is considerable evidence that hydrophobic interactions must play a major role in protein folding. (a) Proteins have hydrophobic cores, implying nonpolar amino acids are driven to be sequestered from water. (b) Model compound studies show 1–2 kcal/mol for transferring a hydrophobic side chain from water into oil-like media, and there are many of them. (c) Proteins are readily denatured in nonpolar solvents. (d) Sequences that are jumbled and retain only their correct hydrophobic and polar patterning fold to their expected native states, in the absence of efforts to design packing, charges, or hydrogen bonding. Hydrophobic and polar patterning also appears to be a key to encoding of amyloid-like fibril structures. What stabilizes secondary structures? Before any protein structure was known, Linus Pauling and colleagues inferred from hydrogen-bonding models that proteins might have α-helices. However, secondary structures are seldom stable on their own in solution. Although different amino acids have different energetic propensities to be in secondary structures, there are also many "chameleon" sequences in natural proteins, which are peptide segments that can assume either helical or β conformations depending on their tertiary context. Studies of lattice models and tube models have shown that secondary structures in proteins are substantially stabilized by the chain compactness, an indirect consequence of the hydrophobic force to collapse. Like airport security lines, helical and sheet configurations are the only regular ways to pack a linear chain (of people or monomers) into a tight space.

COMPUTATIONAL PROTEIN STRUCTURE PREDICTION IS INCREASINGLY SUCCESSFUL

A major goal of computational biology has been to predict a protein's three-dimensional native structure from its amino acid sequence. This could help to (a) accelerate drug discovery by replacing slow, expensive structural biology experiments with faster, cheaper computer simulations, and (b) annotate protein function from genome sequences. With the rapid growth of experimentally determined structures available in the Protein Databank (PDB), protein structure prediction has become as much a problem of inference and machine learning as it is of protein physics. Among the earliest uses of protein databases to infer protein structures were secondary structure prediction algorithms. In the mid-1980s, several groups began using the methods of computational physics atomic force fields plus Monte Carlo sampling to compute the structures of the Metenkephalin, a five-residue peptide. The early 1990s saw significant strides in using databases and homology detection algorithms to assemble structures from homologous sequences and to

recognize folds by threading unknown sequences onto three-dimensional structures from a database. A key advance was the exploitation of evolutionary relationships among sequences through the development of robust sequence alignment methods.

ARE THERE MECHANISMS OF PROTEIN FOLDING?

In 1968, Cyrus Levinthal first noted the puzzle that even though they have vast conformational spaces, proteins can search and converge quickly to native states, sometimes in microseconds. How do proteins find their native states so quickly? It was postulated that if we understood the physical mechanism of protein folding, it could lead to fast computer algorithms to predict native structures from their amino acid sequences. In its description of the 125 most important unsolved problems in science, Science magazine framed the problem this way: “Can we predict how proteins will fold? Out of a near infinitude of possible ways to fold, a protein picks one in just tens of microseconds. The same task takes 30 years of computer time.

The following questions of principle have driven the field: How can all the denatured molecules in a beaker find the same native structure, starting from different conformations? What conformations are not searched? Is folding hierarchical? Which comes first: secondary or tertiary structure? Does the protein collapse to compact structures before structure formation, or before the rate-limiting step (RLS), or are they concurrent? Are there folding nuclei? Several models have emerged. In the diffusion-collision model, microdomain structures form first and then diffuse and collide to form larger structures. The nucleation-condensation mechanism proposes that a diffuse transition state ensemble (TSE) with some secondary structure nucleates tertiary contacts. Some proteins, such as helical bundles, appear to follow a hierarchical diffusion-collision model in which secondary structure forms and assembles in a hierarchical order. In hierarchic condensation, the chain searches for compact, contiguous structured units, which are then assembled into the folded state. Or, proteins may fold via the stepwise assembly of structural subunits called foldons, or they may search for topomers, which are largely unfolded states that have native-like topologies. These models are not mutually exclusive.

****