COMP 598
Advanced Computational Biology Methods & Research

Introduction

Jérôme Waldispühl
School of Computer Science
McGill University

General informations (1)

Office hours: by appointment

Office: TR3018

Contact: jerome.waldispuhl@mcgill.ca

Web: Go to “My Course”

General informations (2)

Evaluation:

- 2 assignments (15% each)
- 2 paper reports & presentations (10% each)
- 1 project (45%)
- Participation (5%)

General informations (3)

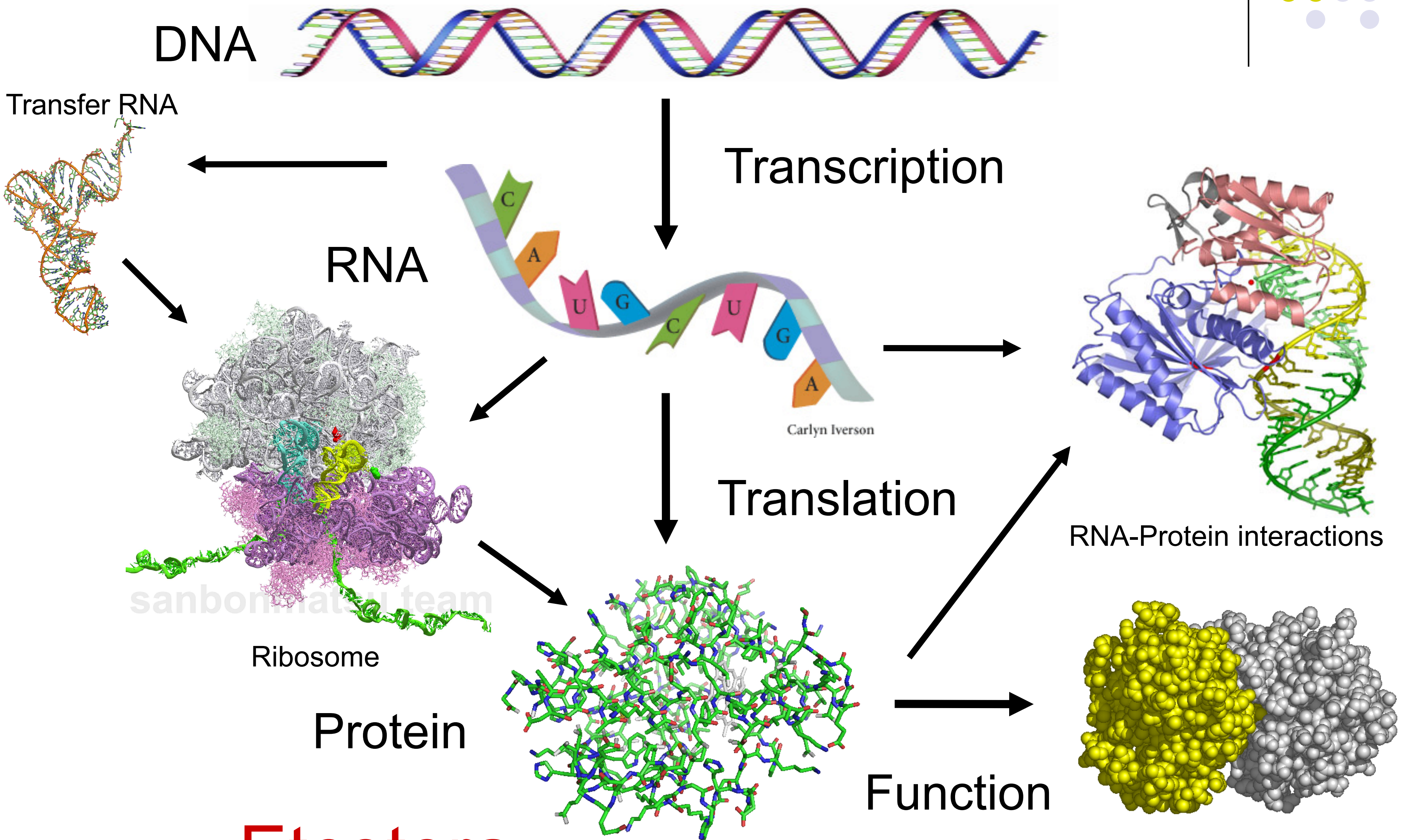
Objective: Extends COMP462/561

Topics: Structural Bioinformatics & System Biology

Background: Algorithmic, Programming & Basic knowledge in Molecular Biology

Invited lectures

Central dogma of biology



DNA

Transcription

Transfer RNA

RNA

Ribosome

Translation

Protein

Function

RNA-Protein interactions

Protein-protein interactions

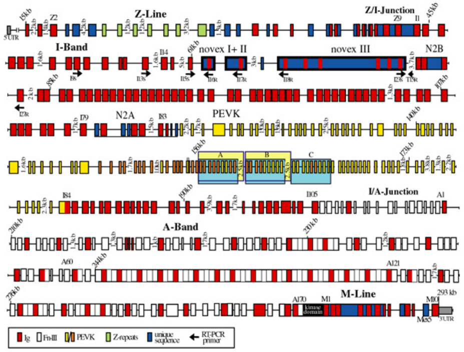
(Images not to scale)

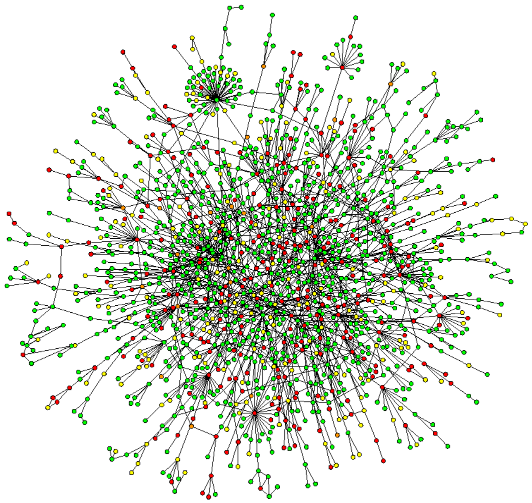
Etcetera…

The 3 components of the Bioinformatics

1. Genomic:

Study of an organism's entire genome.
Huge amount of data, limited to the sequence.



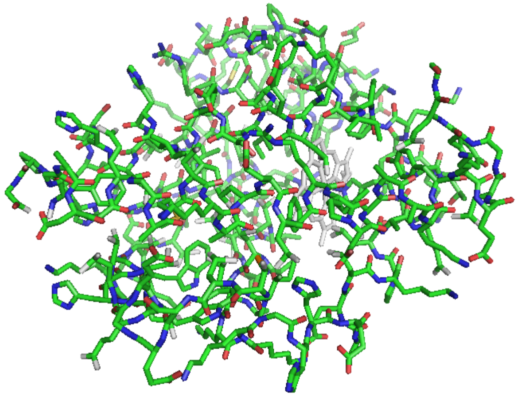


2. System Biology:

Study of complex interactions in biological systems.
High-level of representation, practical interests.

3. Computational Structural Biology:

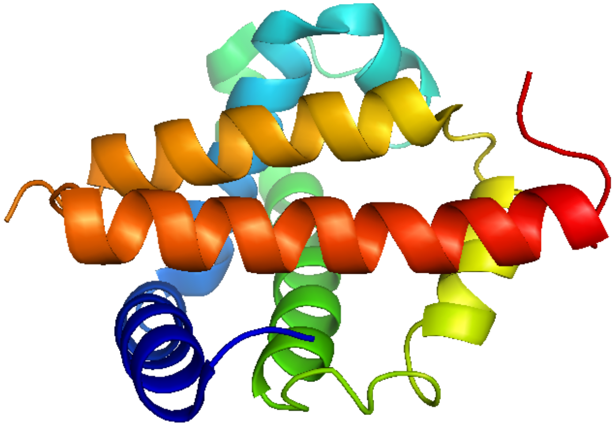
Study of the bio-molecule folding process.
Lack of data in early year of bioinformatics, step toward the function, fill the gap between genomic and system biology.



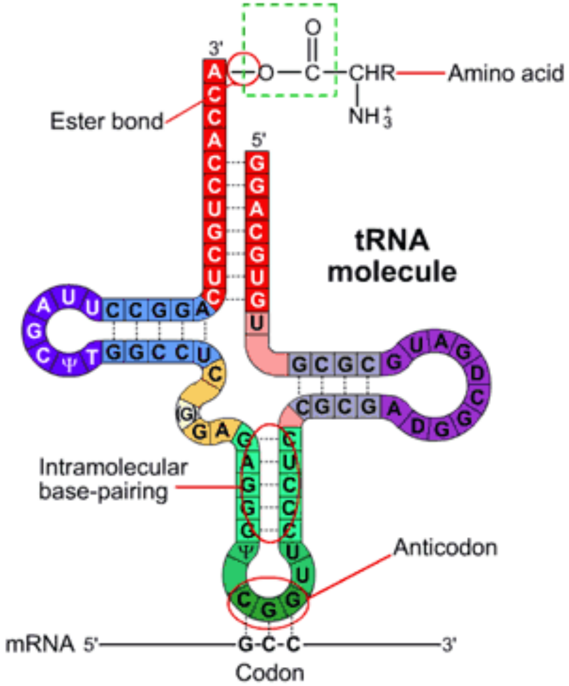
Part 1

Computational Structural Biology

Modeling structures



Protein



RNA

We introduce a intermediate representation (secondary structure) between the sequence (primary) and the 3D structure (tertiary).

Classification of structure & folding prediction methods

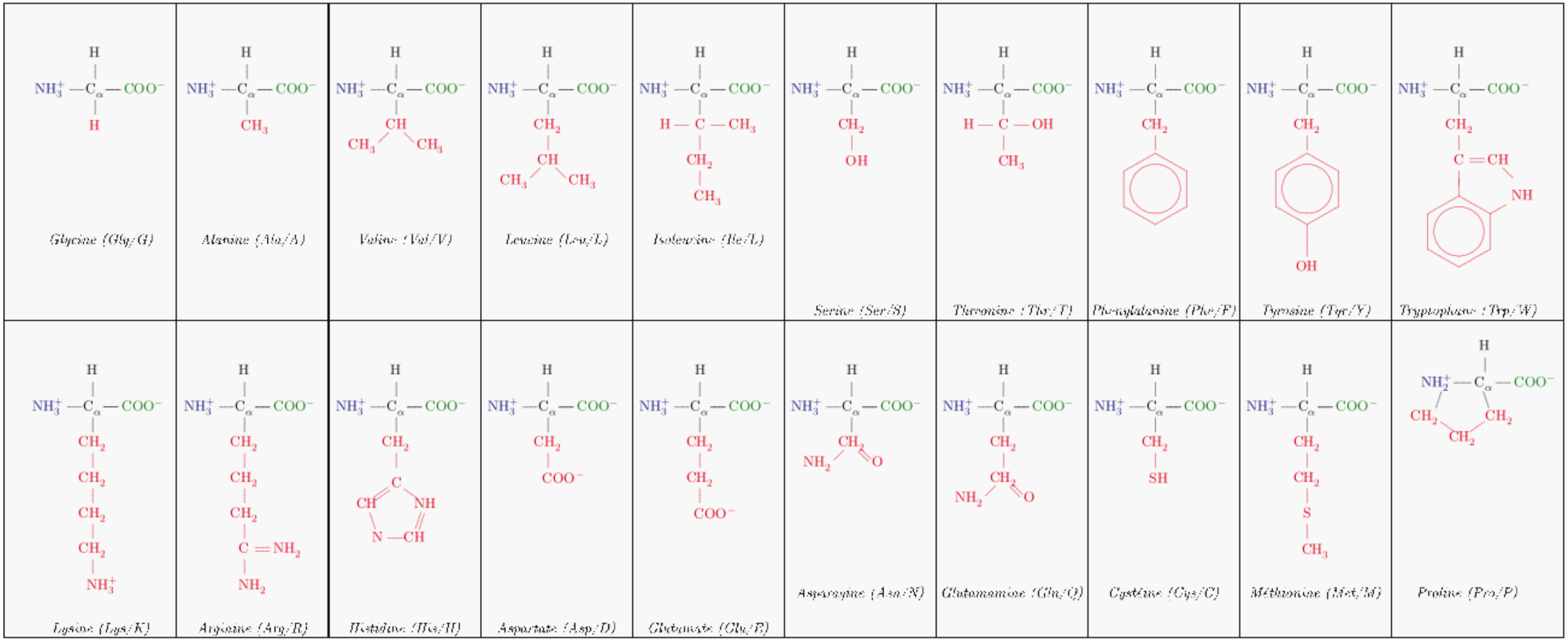
Structure prediction

- **Comparative/Homology modeling:** similar sequences fold the same.
- **Threading/Fold recognition:** fold a sequence on a known 3D template.
- **Ab-initio method:** Sampling the conformational space.

Folding pathway prediction:

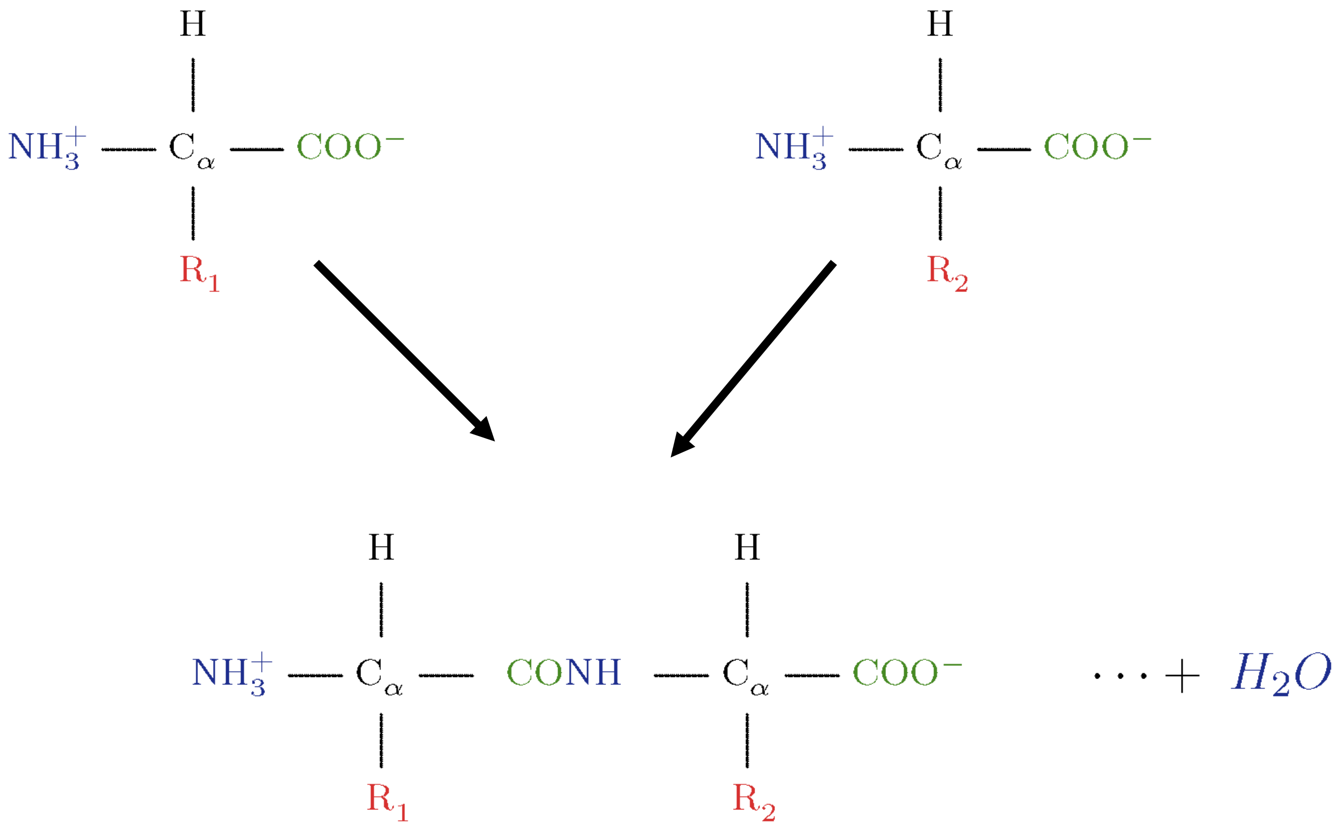
- **Molecular dynamics:** simulation under known laws of physics.
- **Motion planning:** simulation of atomic robotic motions.
- **Coarse grained model:** Discrete modeling of the folding landscape

Protein Structure: amino acids



The 20 amino acids. Building blocks of a protein.
They differs by the nature of their side-chain (radical).

Proteins: Peptide bond

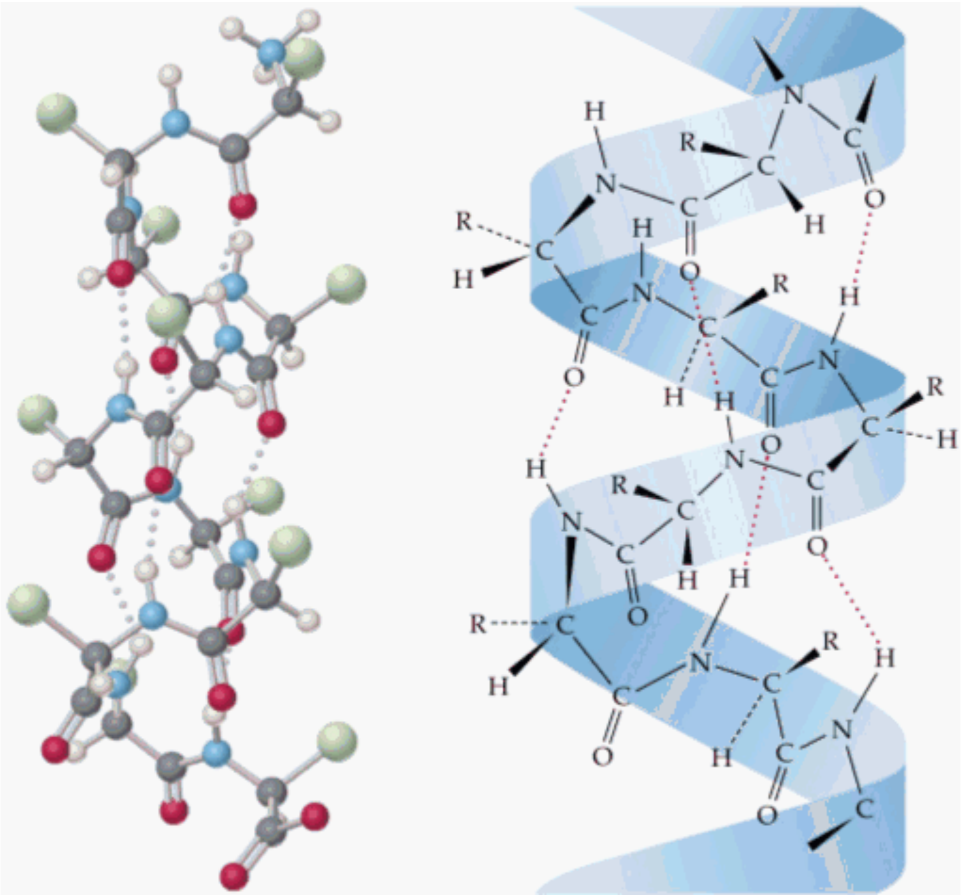


The sequence of amino acids is called the primary structure

Protein secondary structure: α-helices

Features:

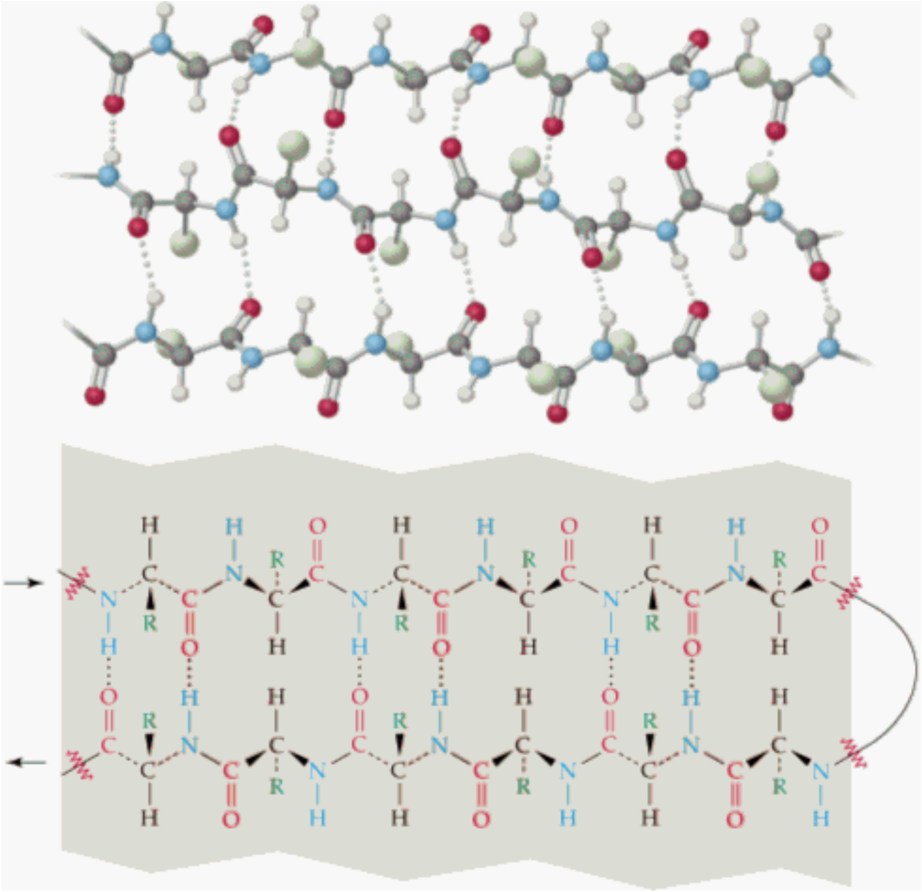
- 3.6 amino acids per turn,
- hydrogen bond between residues n and n+4,
- local motif,
- approximately 40% of the structure.



Protein secondary structure: β-sheets

Features:

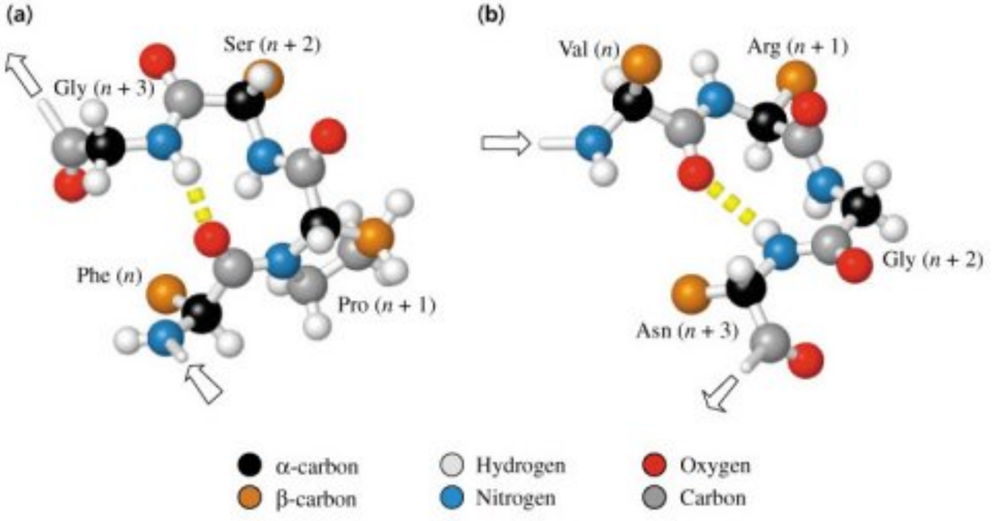
- 2 amino acids per turn,
- hydrogen bond between residues of different strands,
- involve long-range interactions,
- approximately 20% of the structure.

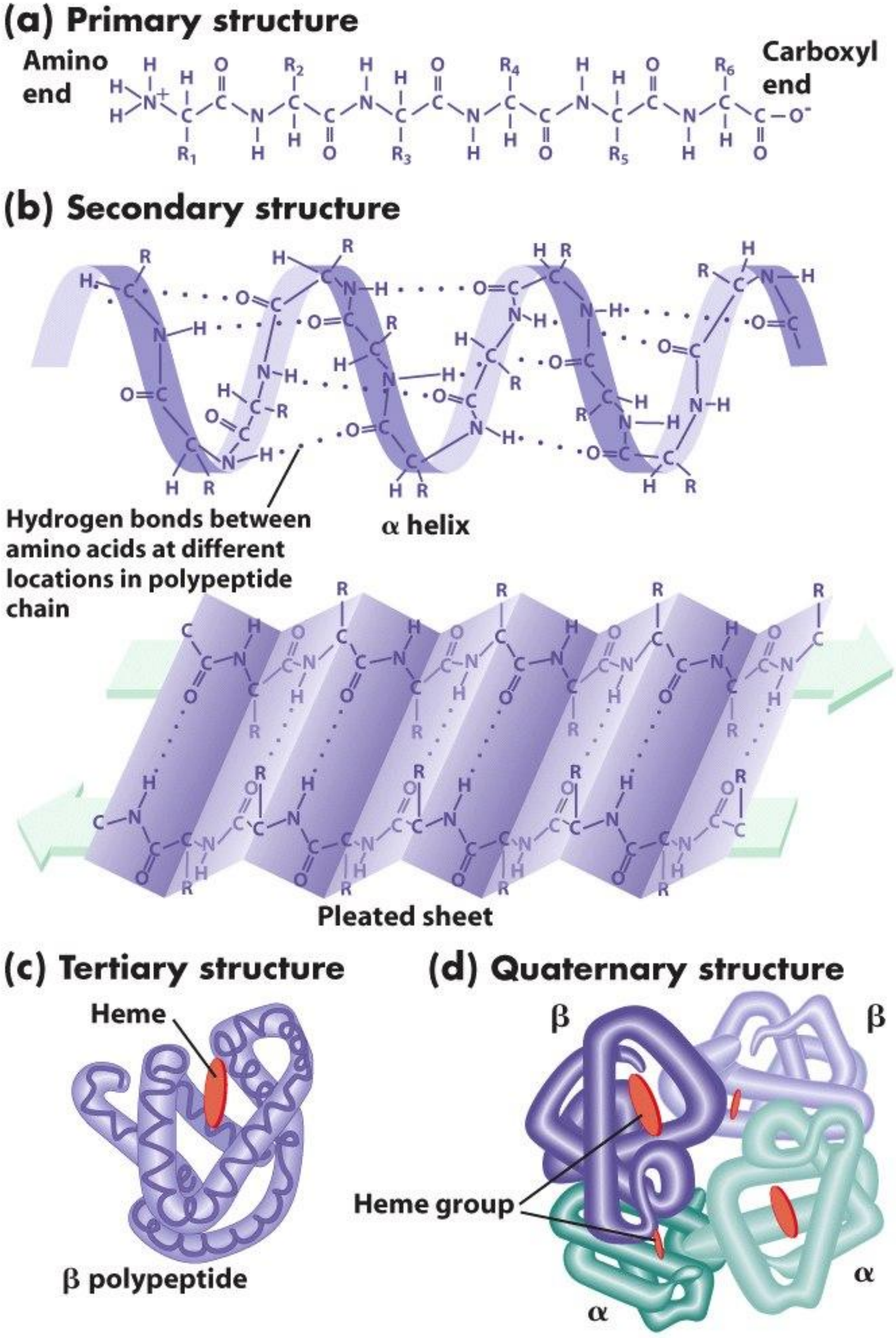


Protein secondary structure: Turns

Features:

- Up to 5 residue length,
- hydrogen bonds depend of type,
- local interactions,
- approximately 5-10% of the structure.

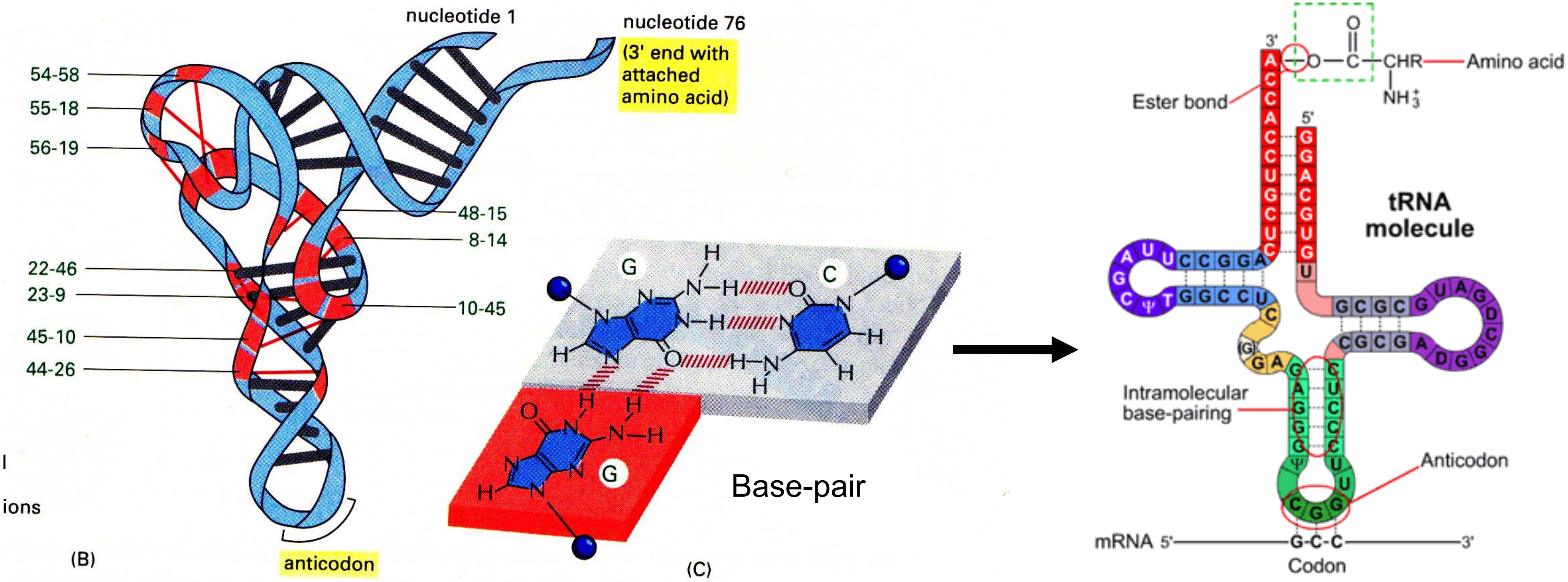




- Secondary structure element are assembled together to form the **tertiary structure**.

- Complexes built from more than one chain form a **quaternary structure**.

RNA structure



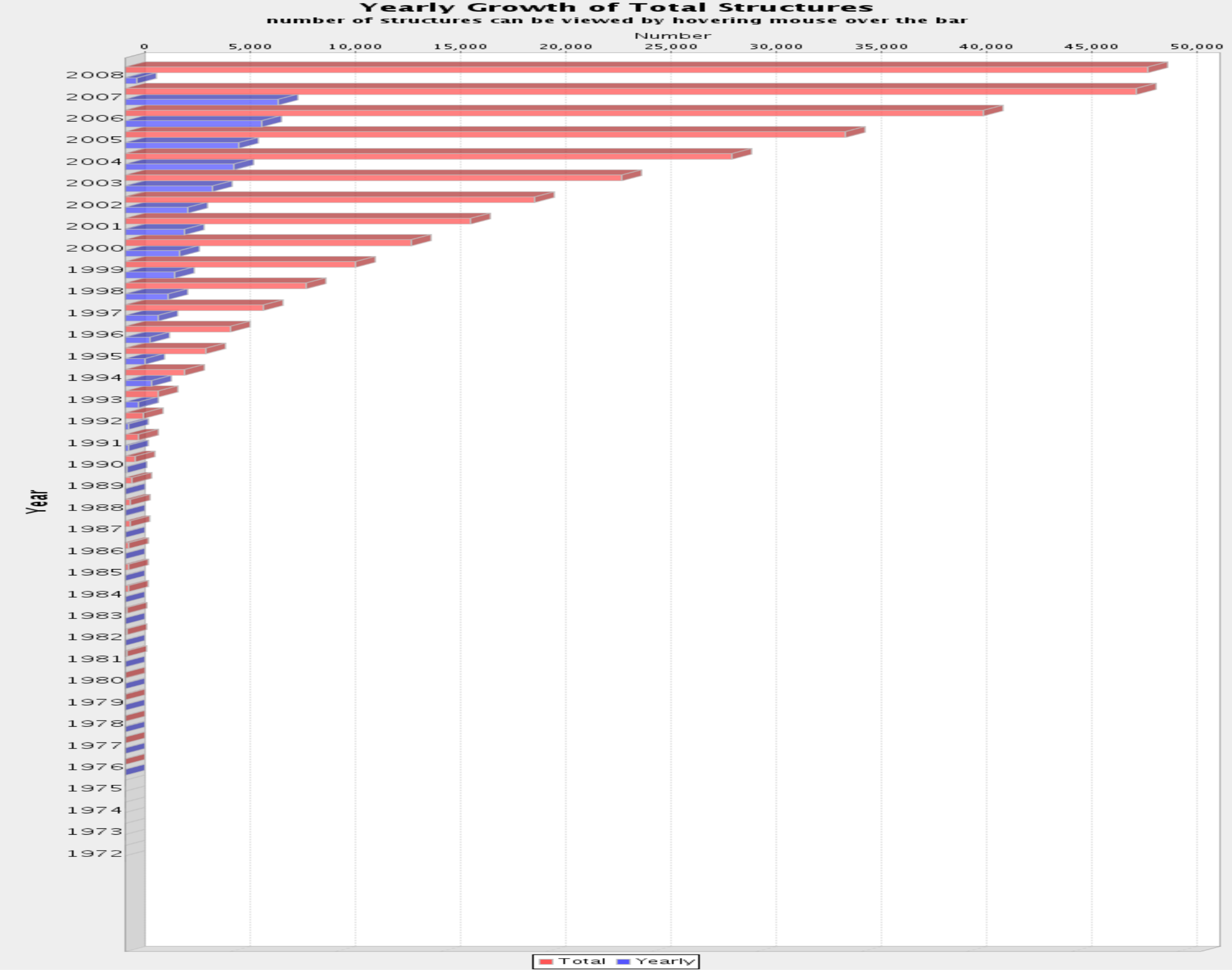
Maximal planar representation (no crossing edges) of the graph of the base-pairs (watson-crick + wobble).

(More details in the next lecture.)

Databases

- Protein Data bank: www.rcsb.org (3D structures)
- MSD-EBI: www.ebi.ac.uk/msd (3D structures)
- PDBj: www.pdbj.org (3D structures)
- UniProtKB/Swiss-Prot: expasy.org/sprot (annotated protein)
- CATH: cathdp.info (structure classification)
- SCOP: scop.mrc-lmb.cam.ac.uk/scop (structure classification)
- BMRB: www.bmrb.wisc.edu (NMR)
- NDB: ndbserver.rutgers.edu (ARNs)

Protein Data Bank



PDB format

Keywords:

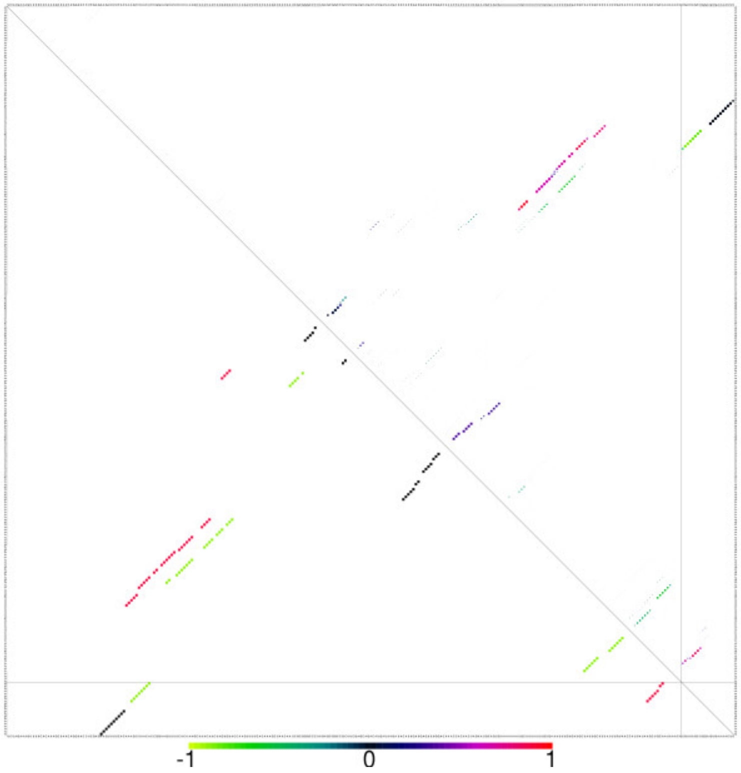
SEQRES: amino acid or nucleic acid sequence.
MODRES: descriptions of modifications to residues.
HELIX: identify the position of helices in the molecule.
SHEET: position of sheets in the molecule.
TURN: identify turns and other short loop turns.
ATOM: atomic coordinates for standard residues.
HETATM: atomic coordinate of atoms within "non-standard" groups.
CONECT: connectivity between atoms for which coordinates are supplied.
HYDBND: specify hydrogen bonds in the entry.
SSBOND: disulfide bond.

PDB format (2)

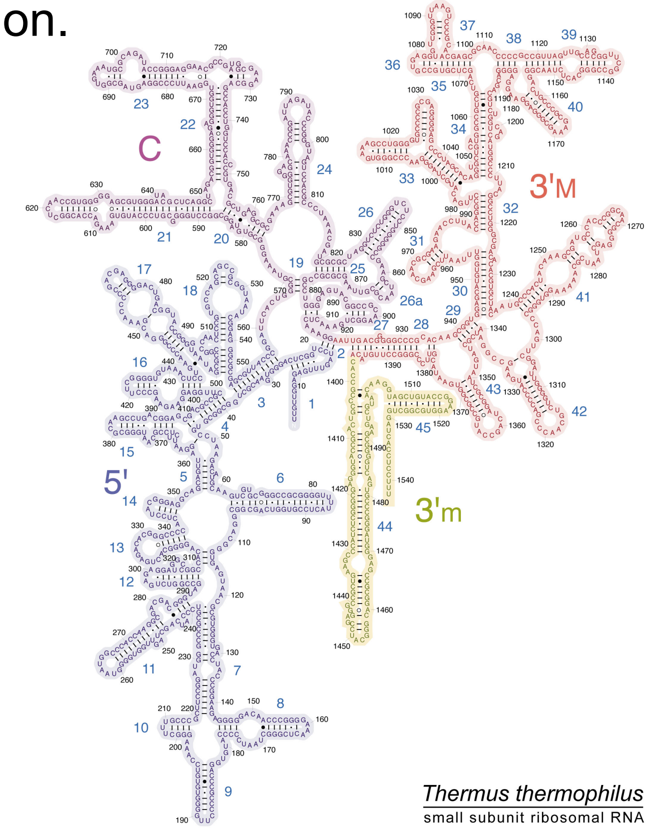
COLUMNS	DATATYPE	FIELD	DEFINITION
1- 6	Record name	"ATOM "	
7-11	Integer	serial	Atom serial number.
13-16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	Char	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
73 - 76	LString(4)	segID	Segment identifier, left-justified.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

Classical secondary structure prediction algorithms.

Lecture 2: Classical secondary structure prediction algorithms.
Lecture 3: RNA sequence/structure alignment.
Lecture 4: Stochastic secondary structure prediction.



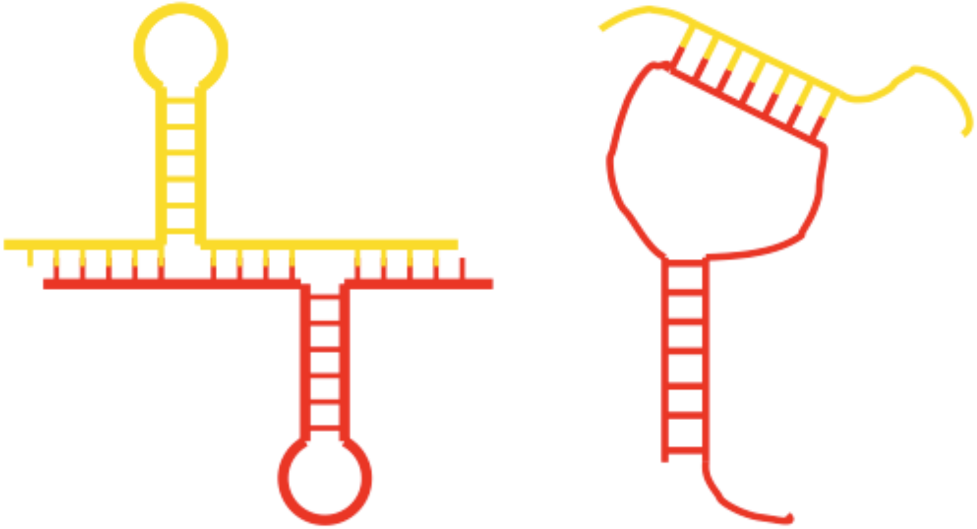
RNA dotplot



Thermus thermophilus
small subunit ribosomal RNA

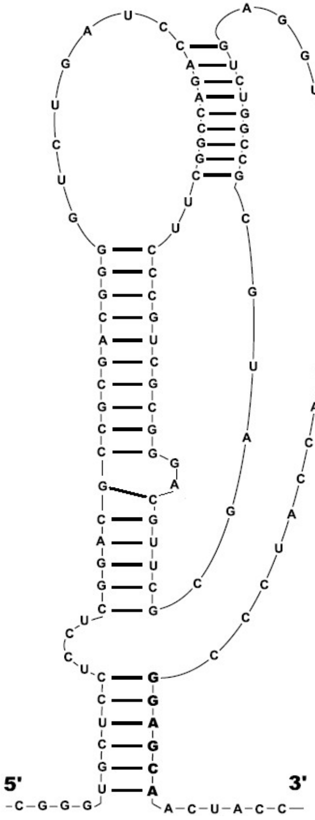
Extended secondary structures

Lecture 5: RNA saturated secondary structures and RNA shapes.
Lecture 6: RNA secondary structures with pseudoknots, RNA-RNA interaction.



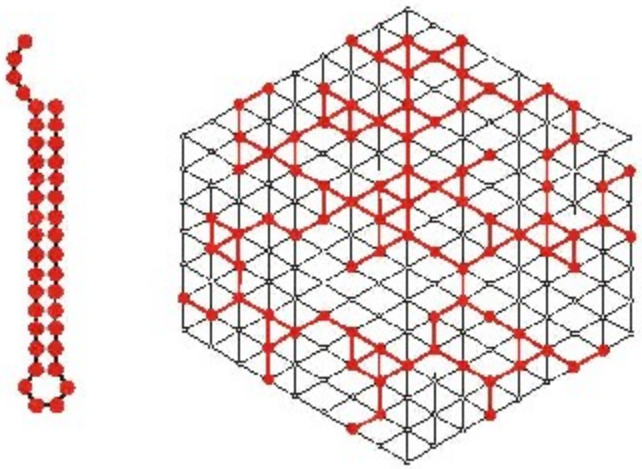
RNA-RNA interaction

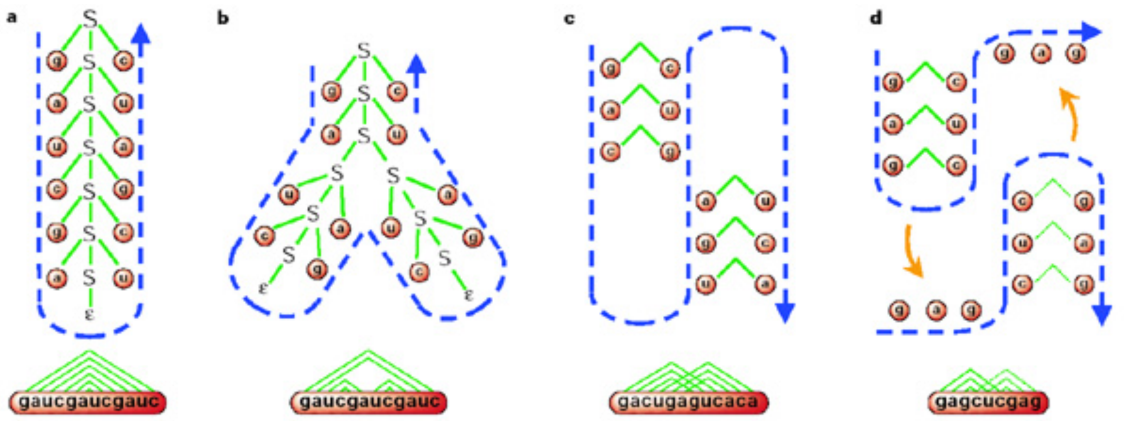
Pseudo-knotted RNA secondary structure:



Lecture 7-9: Theoretical studies in the RNA secondary structure model

Lecture 7: Grammatical modeling of RNA structures.
Lecture 8: Asymptotics of RNA secondary structures
Lecture 9: Evolution, neutral network.
Lecture 10: Synthetic Biology, RNA design.



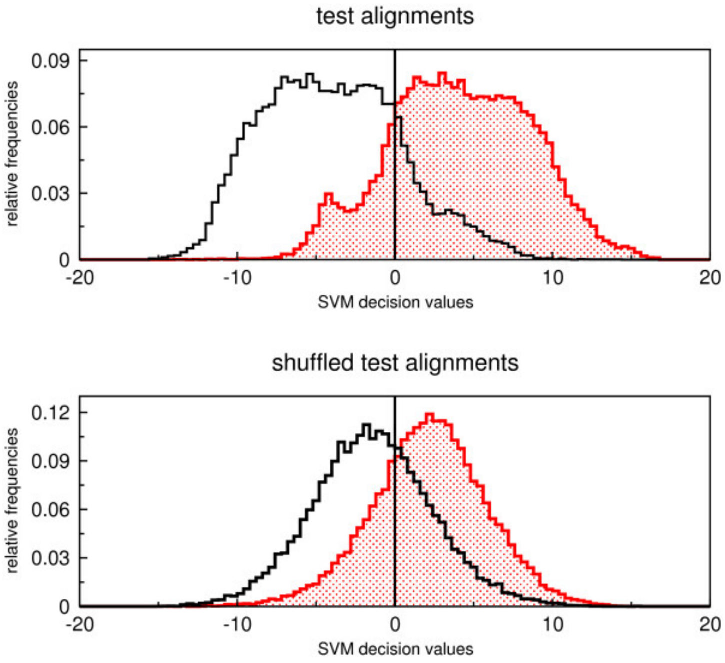


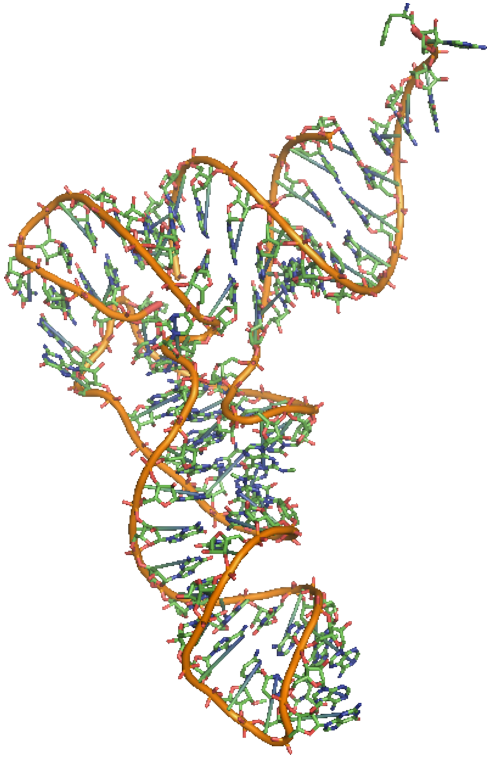
Grammatical modeling of RNA structure

Connected neutral network

Lecture 11-13: Advanced topics

Lecture 11: RNA 3D structure modeling, alignment and prediction.
Lecture 12: Genomic identification of structural RNAs
Lecture 13: RNA folding kynetics

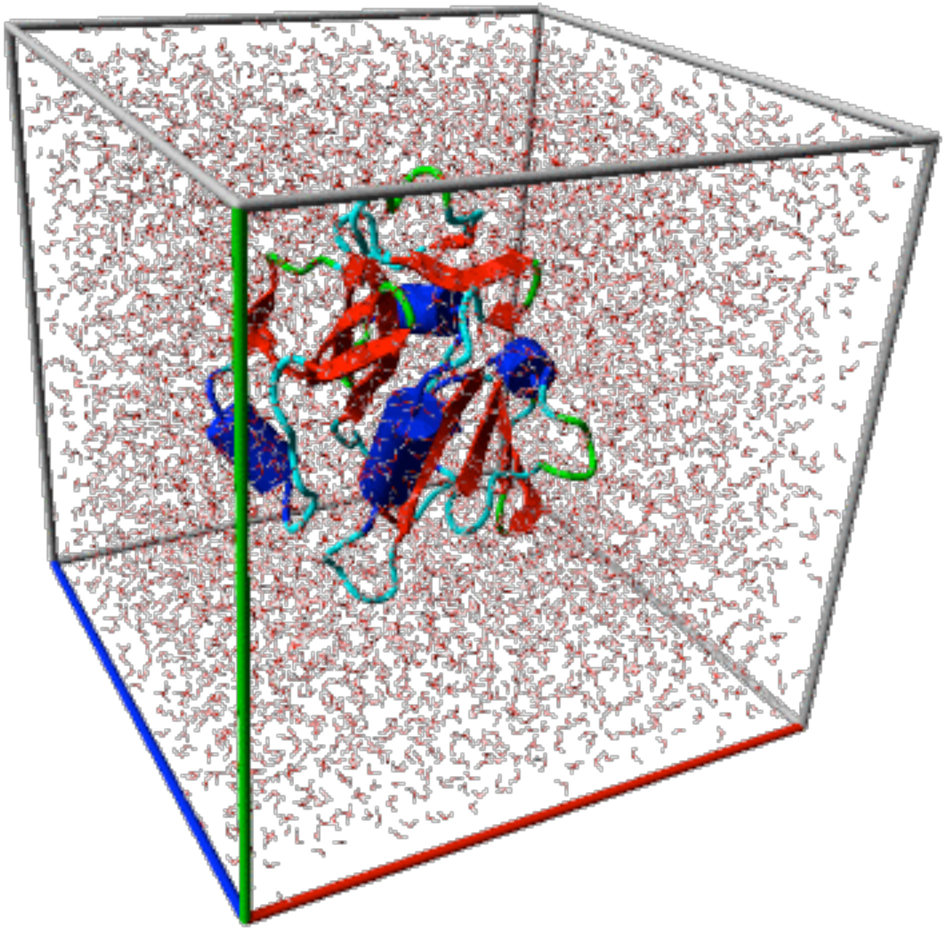




Lecture 14-15: 3D modeling and simulation

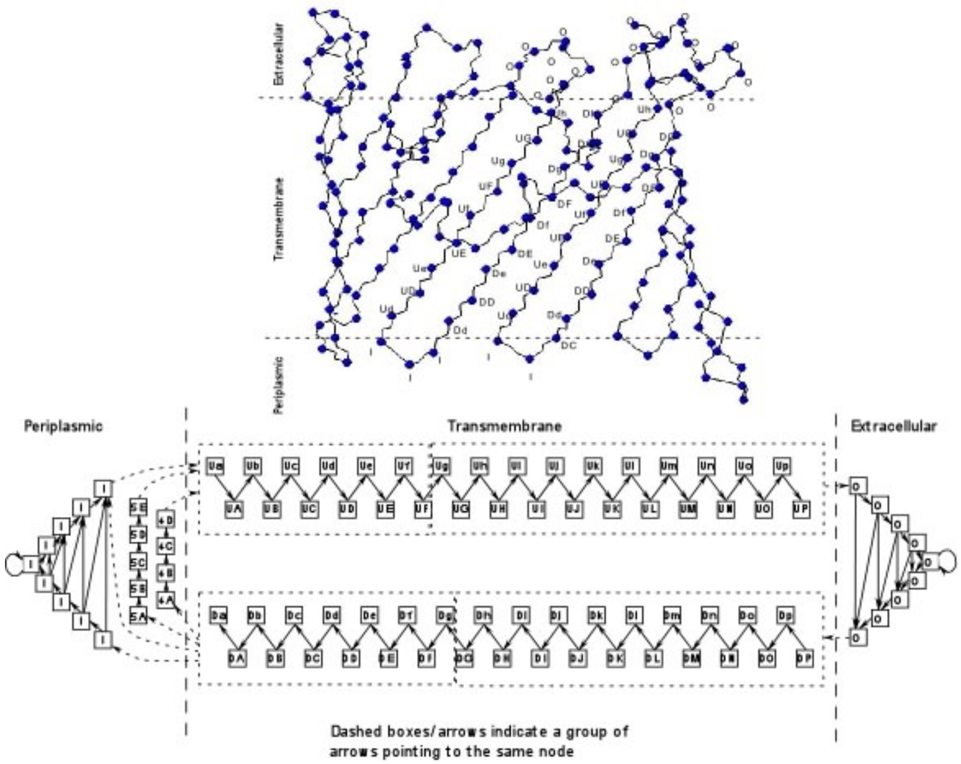
Lecture 14:
Introduction to protein structure prediction.
&
Conformational search and Molecular Dynamics.

Lecture 15:
Threading, fragment assembly, side-chain packing.



Lecture 16-18: template based predictions

Lecture 16: Protein secondary structure prediction.
Lecture 17: Language theory as a tool for protein structure modeling and prediction.
Lecture 18: Transmembrane proteins.

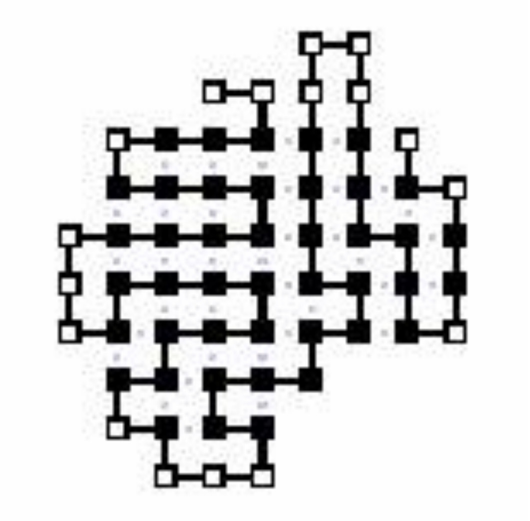


Dashed boxes/arrows indicate a group of arrows pointing to the same node

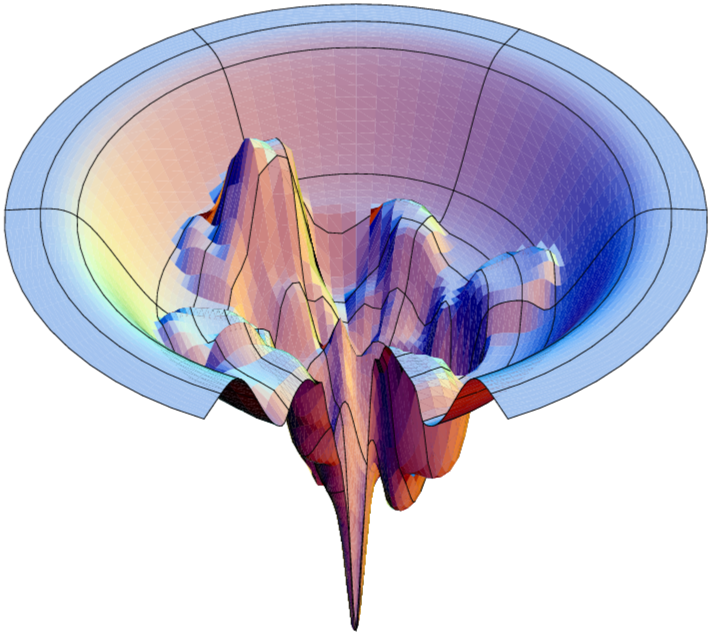
HMM modeling of transmembrane beta-barrel (Bigelow et al., 2010)

Lecture 19-21: Folding pathways

Lecture 19: Protein folding on a lattice models.
Lecture 20: Residue contact prediction & folding pathways.
Lecture 21: Integrative methods.



Protein folding in HP model

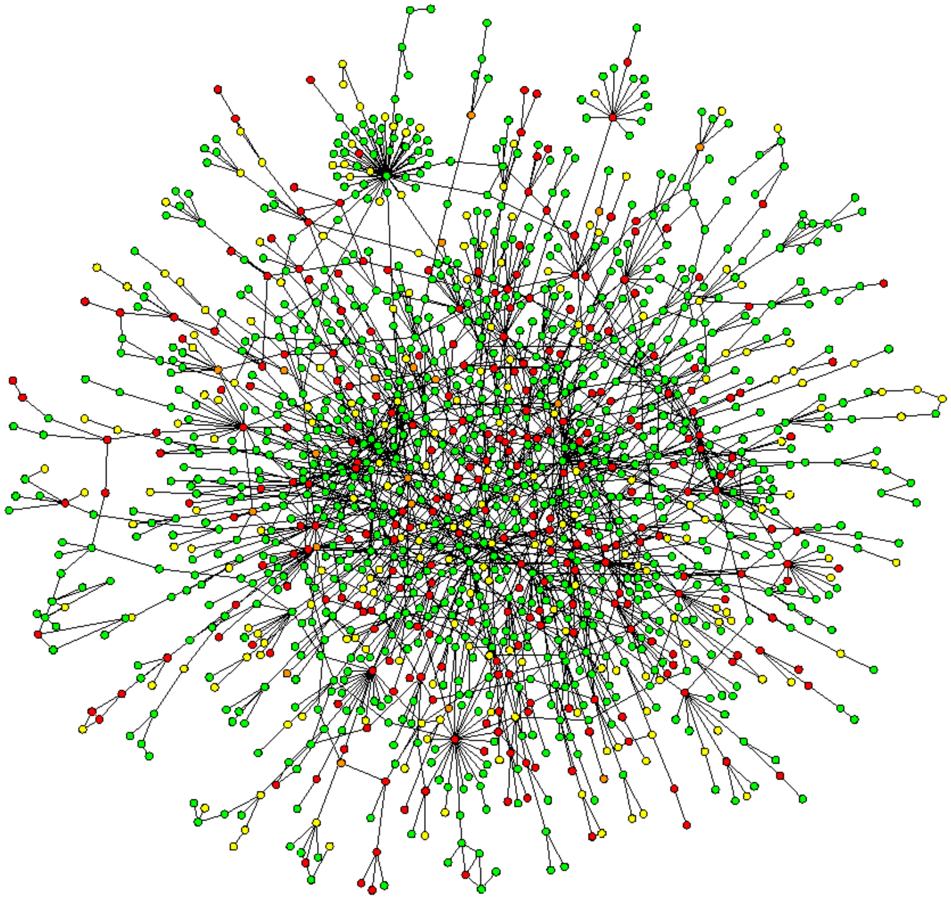


Folding landscape

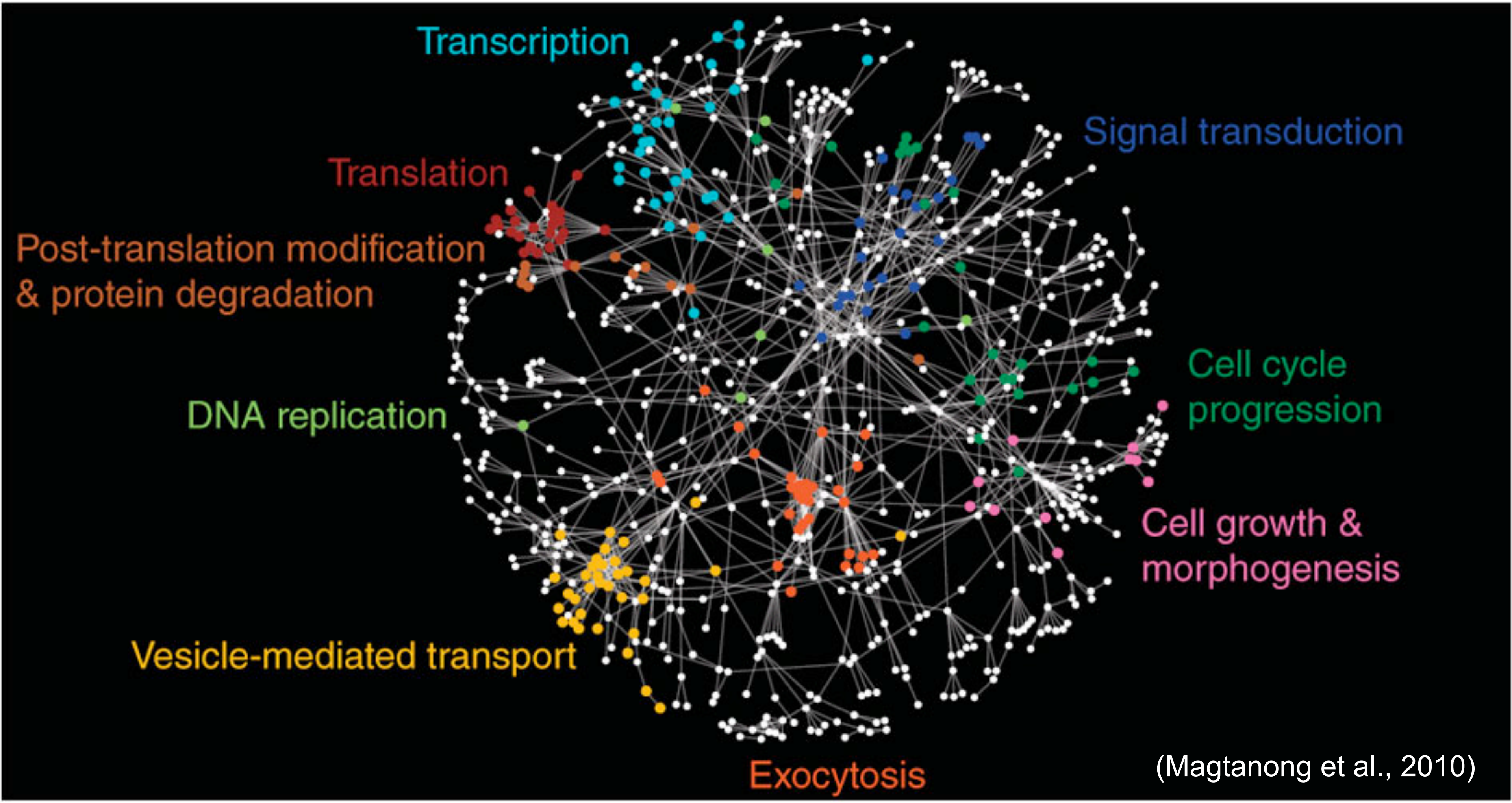
Part 1

System Biology

Protein-protein interaction networks



Gene interaction network

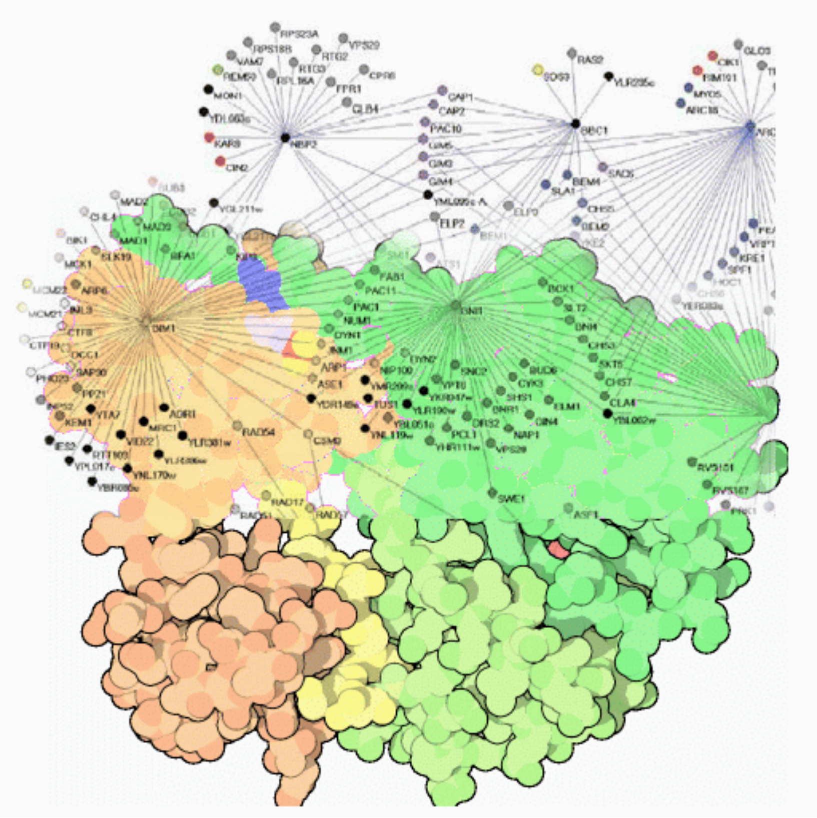

Transcription
Signal transduction
Translation
Post-translation modification & protein degradation
Cell cycle progression
DNA replication
Cell growth & morphogenesis
Vesicle-mediated transport
Exocytosis

(Magtanong et al., 2010)

Lecture 22-23: Algorithms for interaction network

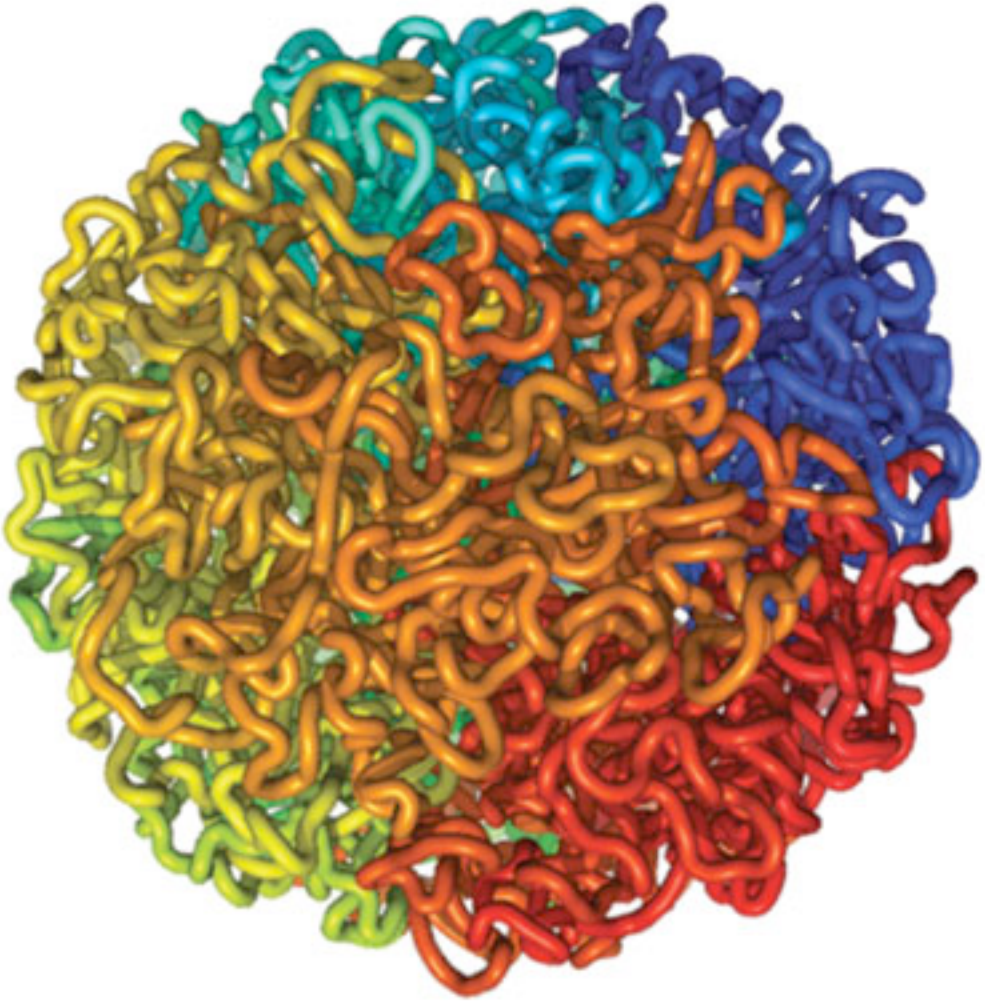
Lecture 22: Modeling interaction network
Lecture 23: Networks alignments & evolution

.



IsoRank (Singh et al., 2008)

Lecture 24: Unifying Structural & System Biology



(Lieberman-Aiden et al., 2009)