Protein structure prediction and more

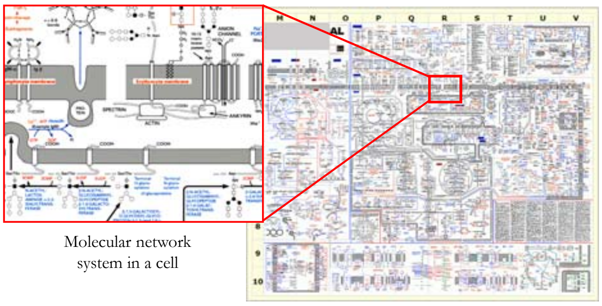
Torgeir R. Hvidsten

This lecture

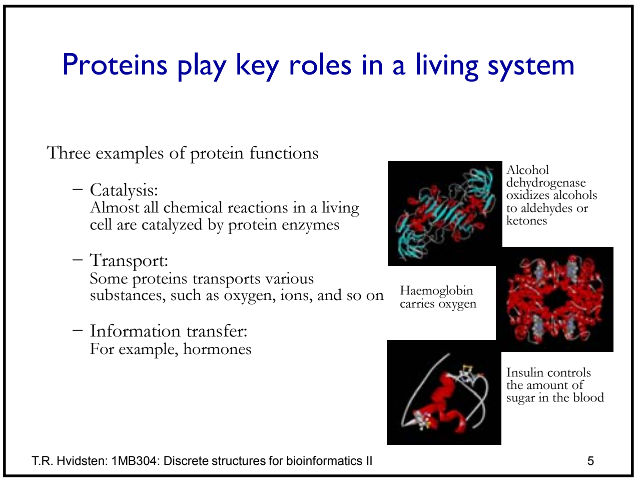
- ➢Protein structure prediction
- ➢The project
- ➢Course summary

Protein structure prediction

Molecular network systems



Molecular network system in a cell



Proteins play key roles in a living system

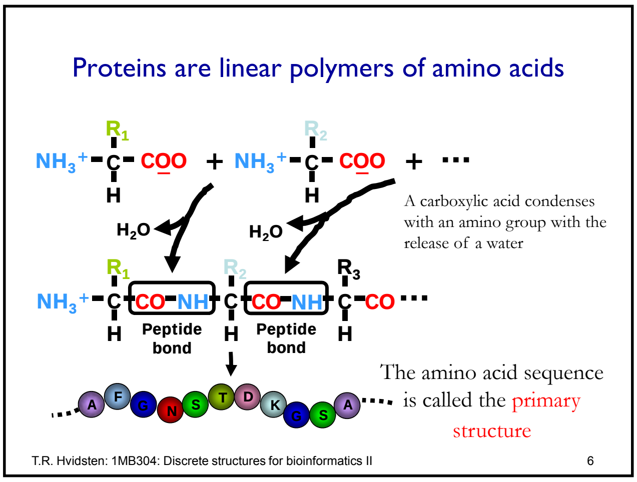
Three examples of protein functions

- Catalysis:
  Almost all chemical reactions in a living cell are catalyzed by protein enzymes
- Transport:
  Some proteins transports various substances, such as oxygen, ions, and so on
- Information transfer:
  For example, hormones

Alcohol dehydrogenase oxidizes alcohols to aldehydes or ketones

Haemoglobin carries oxygen

Insulin controls the amount of sugar in the blood

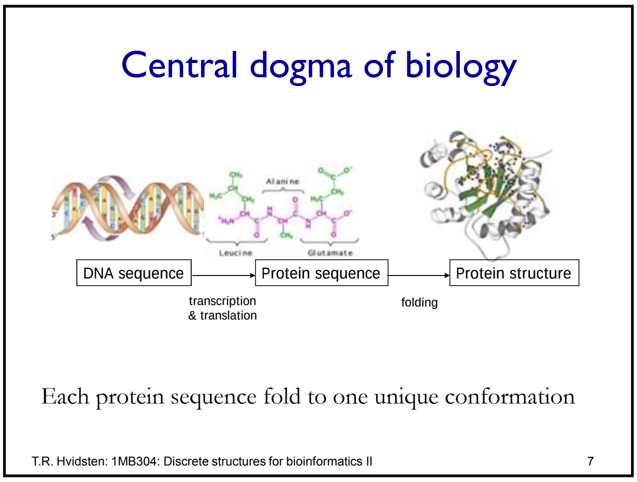


Proteins are linear polymers of amino acids

R_1
NH_3^+ – C – COO + NH_3^+ – C – COO + ...
R_2
H
H_2O
H_2O
R_1
R_2
R_3
NH_3^+ – C – CO – NH – C – CO – NH – C – CO ...
H
Peptide bond
Peptide bond
A
F
G
N
S
T
D
K
G
S
A

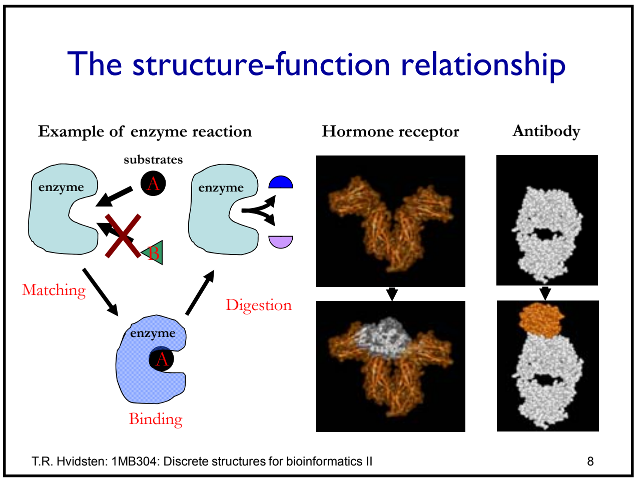
A carboxylic acid condenses with an amino group with the release of a water

The amino acid sequence is called the primary structure

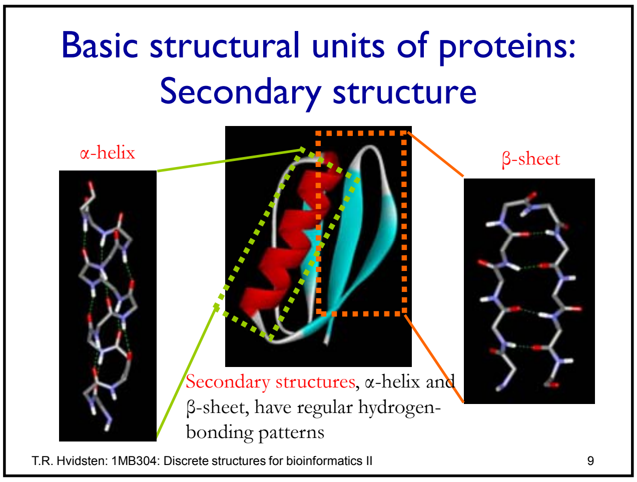


Central dogma of biology

Each protein sequence fold to one unique conformation

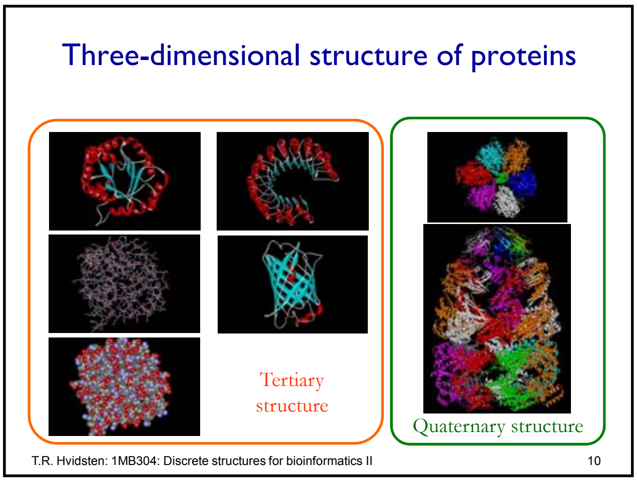


The structure-function relationship

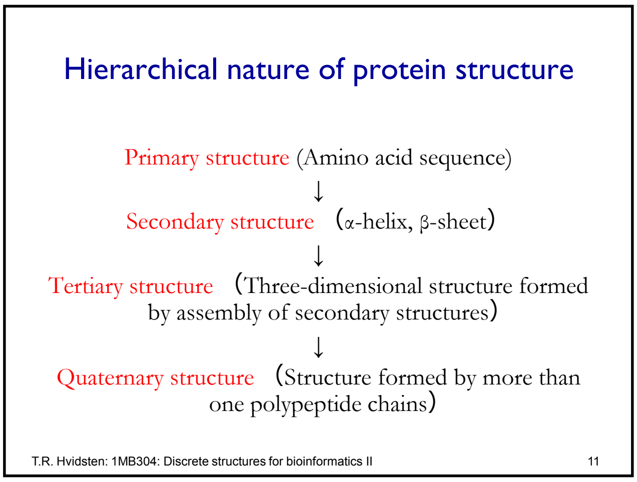


Basic structural units of proteins: Secondary structure

Secondary structures, α-helix and β-sheet, have regular hydrogen-bonding patterns



Three-dimensional structure of proteins



Hierarchical nature of protein structure

Primary structure (Amino acid sequence)

↓

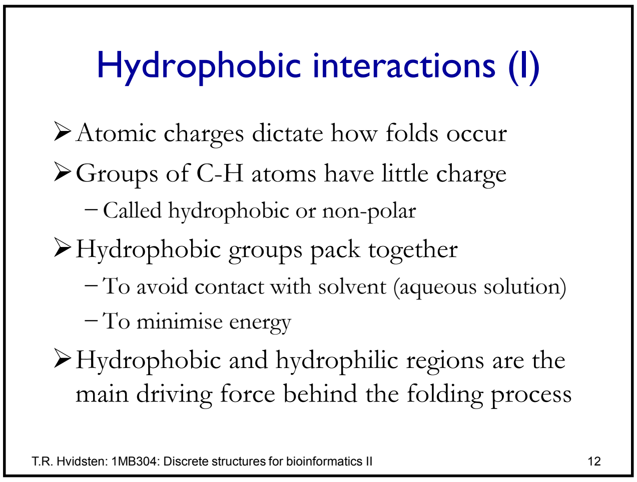
Secondary structure (α-helix, β-sheet)

↓

Tertiary structure (Three-dimensional structure formed by assembly of secondary structures)

↓

Quaternary structure (Structure formed by more than one polypeptide chains)

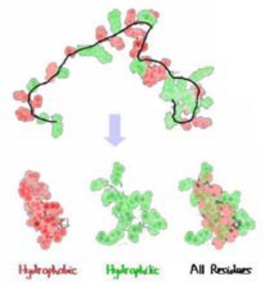


Hydrophobic interactions (I)

- Atomic charges dictate how folds occur
- Groups of C-H atoms have little charge
  - Called hydrophobic or non-polar
- Hydrophobic groups pack together
  - To avoid contact with solvent (aqueous solution)
  - To minimise energy
- Hydrophobic and hydrophilic regions are the main driving force behind the folding process

Hydrophobic interactions (II)

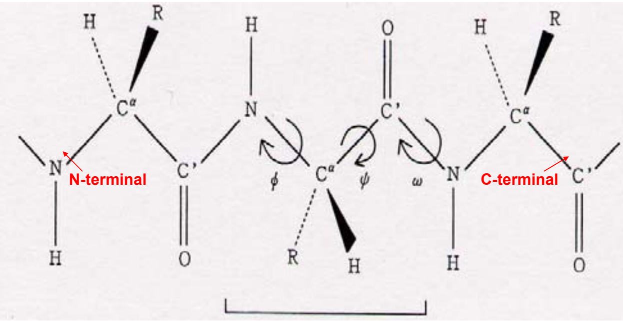
- ➢ Hydrophobicity vs. hydrophilicity
- ➢ Van der Waals interaction
- ➢ Electrostatic interaction
- ➢ Hydrogen bonds
- ➢ Disulfide bonds



Folding is directed mainly by internal residues

- ➢ Mutations that change surface residues are accepted more frequently and are less likely to affect protein conformations than are changes of internal residues
- ➢ This is consistent with the idea of hydrophobic force-driven folding

Structure represented by angels



Protein folding

- ➢ Levinthal's paradox
  - – If for each residue there are only two degrees of freedom (ψ,φ)
  - – Assume each can have only 3 stable values
  - – This leads to 3^{2n} possible conformations
  - – If a protein can explore 10^{13} conformation per second (10 per picosecond)
  - – Still requires an astronomical amount of time to fold a protein
- ➢ This is impossible: protein must fold in a way that does not randomly explore each possible conformations

Molten globule

- Phase 1: Much of the secondary structure that is present in a native proteins forms within a few milliseconds
- Phase 2: Hydrophobic collapse into the Molten globule
  - Slightly larger (5-15% in radius) than the native conformation
  - Significant amount of secondary structure formed
  - Side chains are still not ordered/packed
  - Structure fluctuation is much larger - not very thermodynamically stable

Domains: recurrent units of proteins

- The same or similar domains are found in different proteins
- Each domain has a well determined compact structure and performs a specific function
- Proteins evolve through the duplication and domain shuffling

Protein domains can be defined based on:

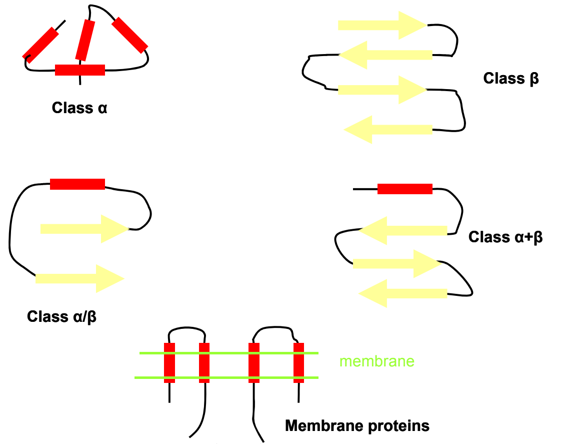
- Geometry: group of residues with a high contact density, number of contacts within domains is higher than the number of contacts between domains
- Kinetics: domain as an independently folding unit
- Physics: domain as a rigid body linked to other domains by flexible linkers
- Genetics: minimal fragment of gene that is capable of performing a specific function

Protein folds

- One domain → one fold
- Fold definition: two folds are similar if they have a similar topology: arrangement/orientation of secondary structure elements (architecture) and connectivity
  - topology = architecture + connectivity
- Fold classification: structural similarity between folds is searched using structure-structure comparison algorithms

Domain/fold classification

- Class α: a bundle of α helices connected by loops on the surface of protein
- Class β: antiparallel β sheets
- Class α/β: mainly parallel β sheets with intervening α helices
- Class α+β: mainly segregated α helices and antiparallel β sheets
- Multidomain proteins: comprise domains representing more than one of the above four classes
- Membrane and cell-surface proteins: α helices (hydrophobic) with a particular length range, traversing a membrane



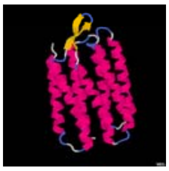
Class α

Class β

Class α/β

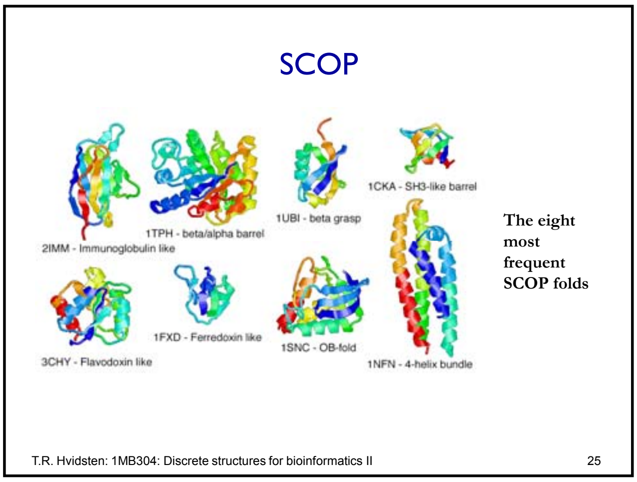
Class α+β

Multi-domain


Membrain-
bound

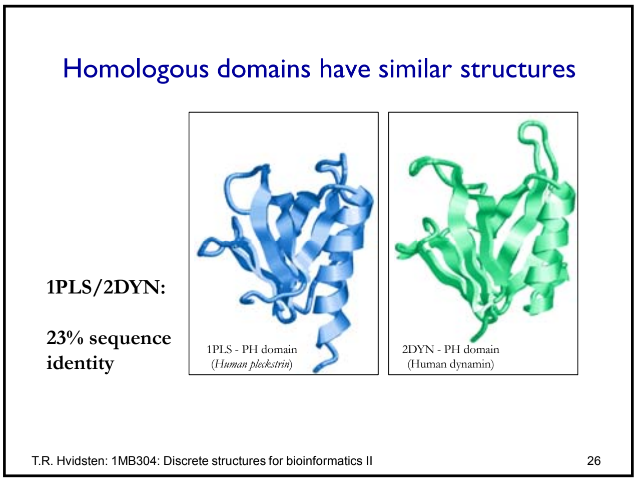
Structural classification of proteins (SCOP)

- The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- Created by manual inspection and aided by automated methods
- Consists of four hierarchical categories:
  - Class, Fold, Superfamily and Family.



SCOP

The eight most frequent SCOP folds



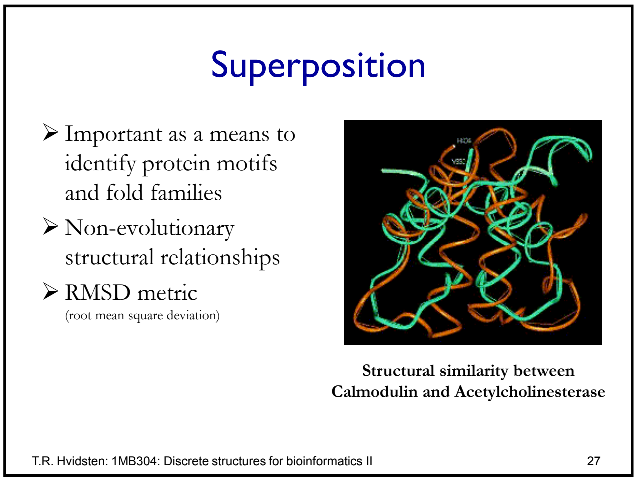
Homologous domains have similar structures

1PLS/2DYN:

23% sequence identity

1PLS - PH domain (*Human pleckstrin*)

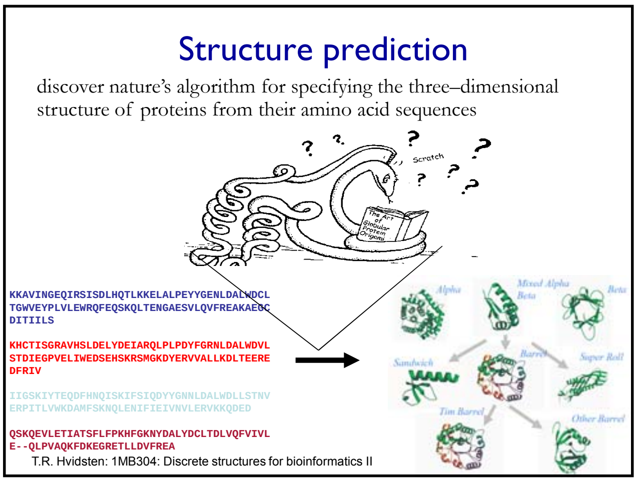
2DYN - PH domain (Human dynamin)



Superposition

- ➢ Important as a means to identify protein motifs and fold families
- ➢ Non-evolutionary structural relationships
- ➢ RMSD metric (root mean square deviation)

Structural similarity between Calmodulin and Acetylcholinesterase



Structure prediction

discover nature's algorithm for specifying the three–dimensional structure of proteins from their amino acid sequences

```
KKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCL
TGWVEYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGC
DITIILS

KHCTISGRAVHSLDELYDEIARQLPLPDYFGRNLDALWDVL
STDIEGPVELIWEDSEHSKRSMGKDYERVVALLKDLTEERE
DFRIV

IIGSKIYTEQDFHNQISKIFSIQDYYGNNLDALWDLLSTNV
ERPITLVWKDAMFSKNQLENIFIEIVNVLERVKKQDED

QSKQEVLETIATSFLFPKHFGKNYDALYDCLTDLVQFVIVL
E--QLPVAQKFDKEGRETLLDVFREA
```

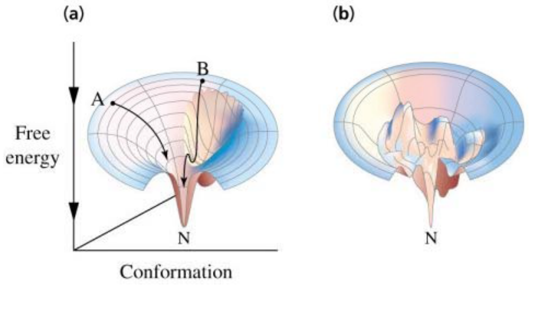
Structure prediction

- Protein structure prediction is the "holy grail" of bioinformatics
- Since structure = function, structure prediction should allow protein design, design of inhibitors, etc
- Huge amounts of genome data - what are the functions of all of these proteins?

Assumptions

- Assumption 1: All the information about the structure of a protein is contained in its sequence of amino acids
- Assumption 2: The structure that a (globular) protein folds into is the structure with the lowest free energy
- Finding native-like conformations require:
  - A scoring function (potential)
  - A search strategy.

The free energy surface of a protein



Physics-based protein simulation

- All quantum mechanics (QM) calculation is not feasible
- QM can be applied to a small set of atoms
  - Modeling of an active site
  - Can get total energies (binding vs. non-binding, pK_a etc.), wave function (charge distribution)
  - QM/MM simulations (i.e. remaining atoms are treated with Molecular Mechanics)

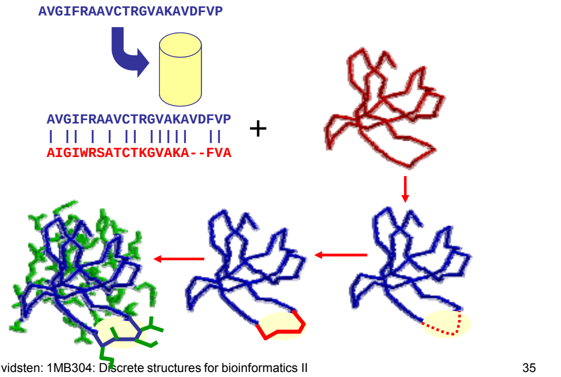
Problems

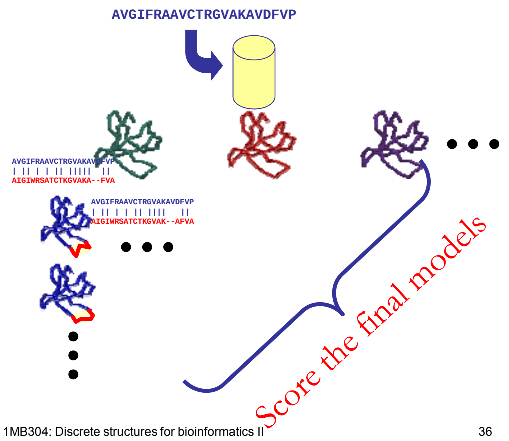
- ➢ Is the energy function correct?
  - – Precise enough to discriminate other non-native structure.
  - – Yet simple enough for computers to carry out efficiently.
- ➢ Is the conformational search good enough to cover the global minimum?
- ➢ Protein folding without any prior knowledge about protein structure is a difficult task.
- ➢ Protein structure prediction is often quoted as an “NP complete problem”, i.e. the complexity of the problem grows exponentially as the number of residues increases

Flavors of “knowledge-based” structure prediction

- ➢ Experimental Methods
  - – X-ray crystallography
  - – NMR spectroscopy
- ➢ Computational methods
  - – Homology/comparative modeling
  - – Fold recognition (threading)
  - – Ab initio (de novo, new folds) methods (Ab initio: “from the beginning”.

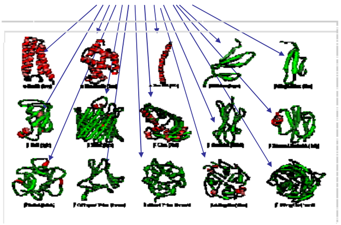
Comparative modeling





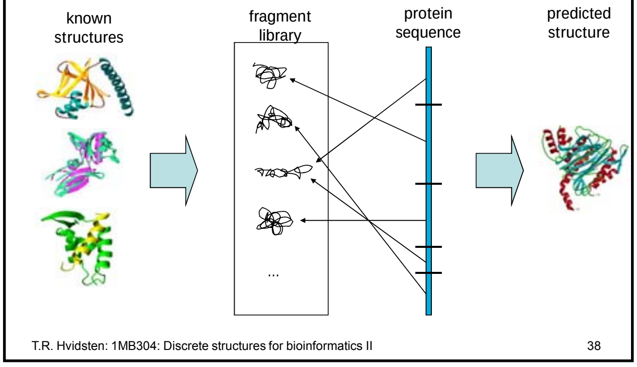
Fold recognition

```
AVGIFRAAVCTRGVAKAVDFVPVESMETTMRSPV
FTDNSSPPAVPQSFQVAHLHAPTGSGKSTKVPAA
YAAQGYKVLVLNPSVAATLGFGAYMSKAHGIDPN
IRTGVRTITTGAPVTYSTYGKFLADGGCSGGAYD
IIICDECHSTDSTTILGIGTVLDQAETAGARLVV
LATATPPGSVTVPHPNIEEVALSNTGEIP
```



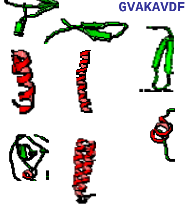
Score and select model

Fragment assembly



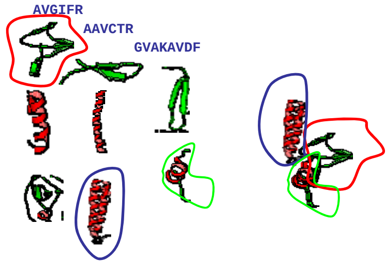
New fold/*ab initio* prediction

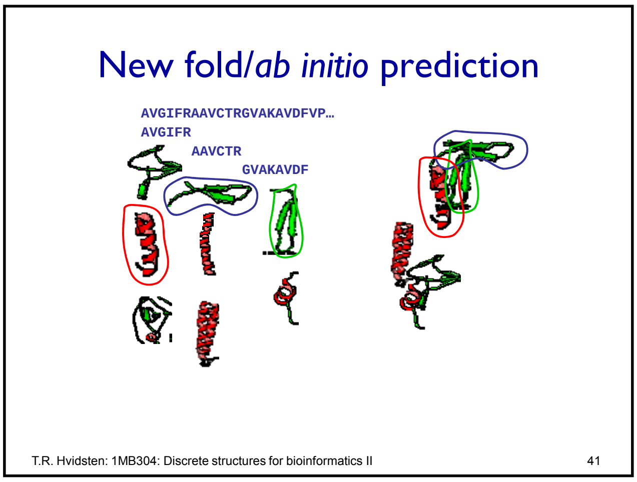
AVGIFRAAVCTRGVAKAVDFVP…
AVGIFR
AAVCTR
GVAKAVDF



New fold/*ab initio* prediction

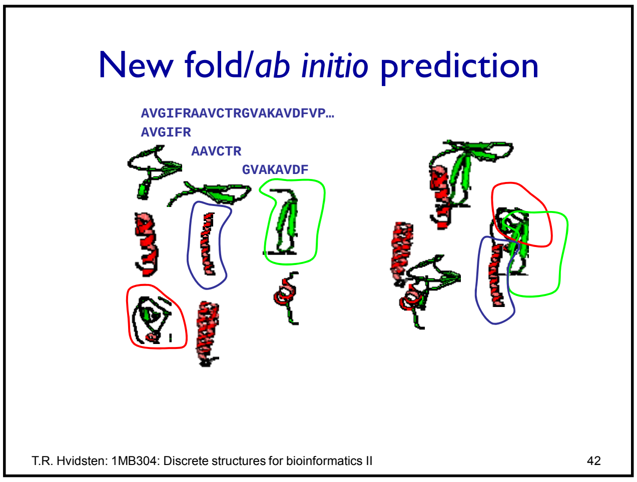
AVGIFRAAVCTRGVAKAVDFVP…
AVGIFR
AAVCTR
GVAKAVDF





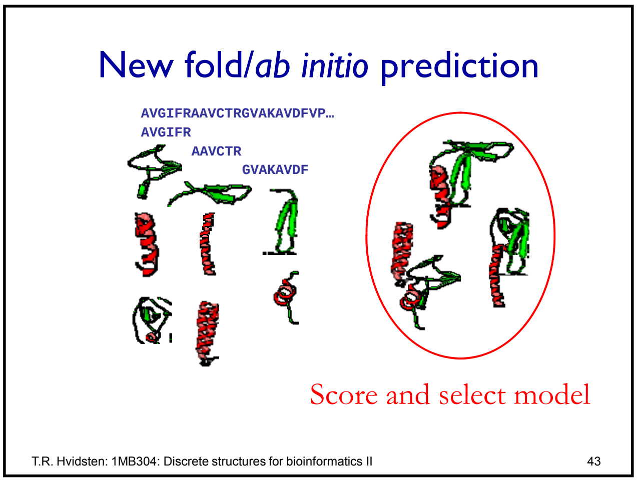
New fold/*ab initio* prediction

AVGIFRAAVCTRGVAKAVDFVP…
AVGIFR
AAVCTR
GVAKAVDF



New fold/*ab initio* prediction

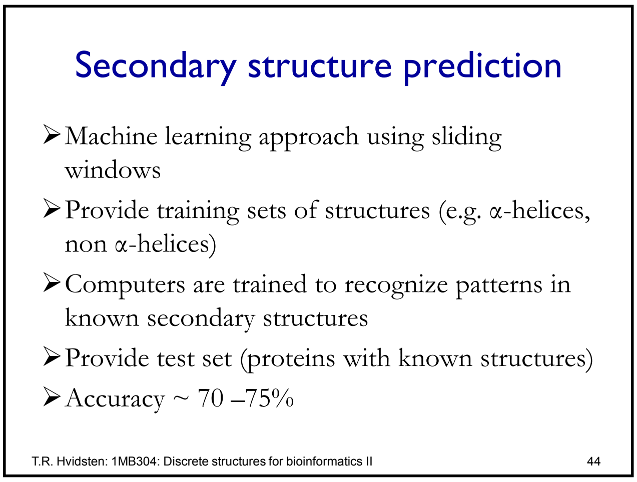
AVGIFRAAVCTRGVAKAVDFVP…
AVGIFR
AAVCTR
GVAKAVDF



New fold/*ab initio* prediction

AVGIFRAAVCTRGVAKAVDFVP…
AVGIFR
AAVCTR
GVAKAVDF

Score and select model

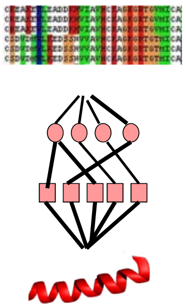


Secondary structure prediction

- Machine learning approach using sliding windows
- Provide training sets of structures (e.g. α-helices, non α-helices)
- Computers are trained to recognize patterns in known secondary structures
- Provide test set (proteins with known structures)
- Accuracy ~ 70 –75%

Example: The PhD algorithm

- ➢ Search databases and select high scoring homologues
- ➢ Create a sequence “profile” from the resulting multiple alignment
- ➢ Input the profile into a trained two-layer neural network to predict the structure and to “clean-up” the prediction



Project (background)

Local descriptors of protein structure

- ➢ A local descriptor of protein structure consists of several short backbone fragments that are close to each other in 3D space but not necessarily on a protein sequence.
- ➢ Possible applications
  - automatic structural classification of proteins
  - detecting particular spatial motifs in proteins
  - identifying boundaries of protein domains
  - pair-wise and multiple structure alignments of proteins
  - protein tertiary structure prediction
    - fold recognition
  - protein function prediction

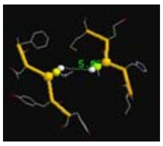
SCOP

A hierarchy according to evolutionary origin and structural similarity

All alpha proteins (α)	[CLASS]
Globin-like	[FOLD]
Globin-like	[SUPER-FAMILY]
Truncated hemoglobin	[FAMILY]
Globins	[FAMILY]
Alpha-helical ferredoxin	[SUPER-FAMILY]
Long Alpha-hairpin	[FOLD]
All beta proteins (β)	
Alpha and beta proteins (α/β)	
Alpha and beta proteins (α+β)	
Multi-domain proteins (alpha and beta)	
Membrane and cell surface proteins and peptides	
…	

Local descriptors of protein structure

Distance definition



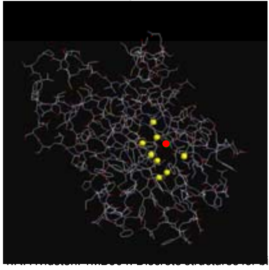
$$\left|C_{\beta x}^{(i)} - C_{\beta x}^{(j)}\right| < 6.5\mathring{A}$$

or

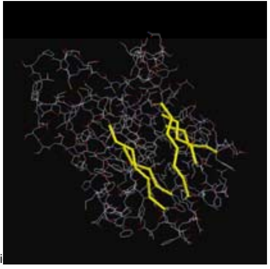
$$6.5\mathring{A} < \left|C_{\beta x}^{(i)} - C_{\beta x}^{(j)}\right| < 8.0\mathring{A}$$

$$\text{and} \left|C_{\beta x}^{(i)} - C_{\beta x}^{(j)}\right| < \left|C_{\alpha}^{(i)} - C_{\alpha}^{(j)}\right| - 0.75\mathring{A}$$

Local neighborhood



Descriptor: 1nuk_A#96



Descriptor group

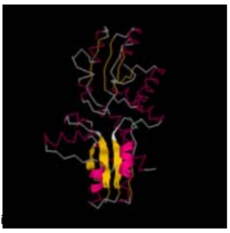
Sequence fragments

Descriptor	Segment 1		Segment 2		Segment 3		Segment 4		Segment 5	
1qgoa_#8	4-10	ALLVVSF	39-43	FRAFT	63-67	LQALQ	77-83	VAIQSLH	91-95	EKIVR
1qo2a_#78	74-80	EHIQIGG	47-51	IHVVD	67-71	EKLSE	97-101	--RRQIV	89-93	EKLRK
1rpxa_#73	69-75	LPLDVHL	40-44	IHVDV	58-62	LVVDS	93-97	--DIVSV	85-89	PDFIK
1nsj__#82	78-84	NAVQLHG	58-62	GVFVN	66-70	EKILD	98-104	ILVIKAV	89-93	ELCRK
1mla_1#7	3-9	QFAFVFP	87-91	MMAGH	262-266	EYMAA	270-276	EHLYEVG	283-287	GLTKR
1qfja2#108	104-110	PMILIAG	134-138	TIYWG	183-187	TAVLQ	195-201	HDIYIAG	207-211	KIARD
1efvb1#8	4-10	LRVLVAV	119-123	LVLLG	47-51	EEAVR	59-65	KEVIAVS	76-80	RTALA
1iow_1#8	4-10	KIAVLLG	38-42	YPVDP	48-52	TQLKS	56-62	QKVFIAL	70-74	GTLQG
1yaca_#57	53-59	PTILTTS	80-84	PYIAR	97-101	VKAVK	14-20	AVLLVDH	120-124	AFPAL
1ig0a2#188	184-190	ISLLALG	40-44	TLLIL	128-132	TKCVN	216-222	FKLCYMT	200-204	VHSIT

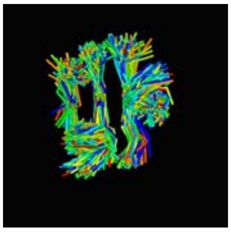
Grouping function

- number of segments
- length of segments
- shape of individual segments
- number of pairs that fit under a specific RMSD cutoff
- overall RMSD score between descriptors

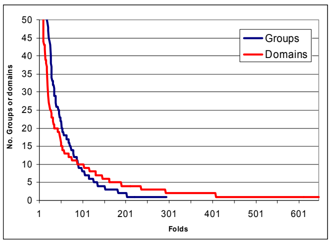
Descriptor: 1qgoa_#8



Structurally similar descriptors



Library of common local structures (I)



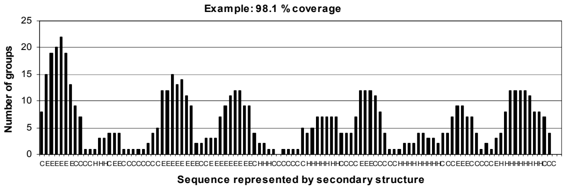
- Training set: 4013 protein domains in ASTRAL 1.57 (less than 40% sequence identity to each other)

▼

- 4084 descriptor groups (fold-oriented) with at least 7 descriptors with at least 3 segments

Library of common local structures (II)

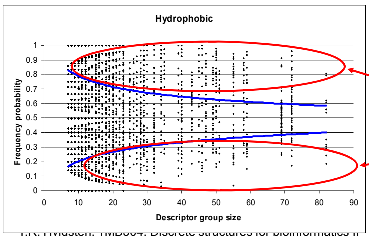
- Coverage: Fraction of a sequence that structurally match at least one descriptor group
  - Domains in training set: 67%
  - Separate domains (< 40% seq. id to training set): 50%



Signal extraction (I)

```
                DESCRIPTOR GROUP
  1ay7_B#81  ITIIL  LVLEW--  SVLQVFREA
 1b9y_C#114  GFVYE  VKFCKIR  ALNSSLEC-
   1bpm#143  VSAKL  PSVEVDP  ---EGAVLG
 1cjm_A#190  VLYLF  VVYVARN  ---QEWWEL
   1cy4_A#5  KALVI  DYVVKSS  SELKQLAEK
 1d4g_A#382  VGVHF  VPAINVN  TGVHNLYKM
  1dos_A#11  PGVIT  VPVILHT  SGAHHVHQM
```

- ➢ Signal extraction aims at identifying sequence-derived patterns in groups
- ➢ Based on observed frequencies of amino acids and amino acid substitution groups in specific positions



Hydrophobic amino acids

Alanine (A)
Glycine (G)
Isoleucine (I)
Leucine (L)
Methionine (M)
Phenylalanine (F)
Proline (P)
Tryptophan (W)
Valine (V)

37% significant

Signal extraction (II)

```
                DESCRIPTOR GROUP
  1ay7_B#81  ITIIL  LVLEW--  SVLQVFREA
 1b9y_C#114  GFVYE  VKFCKIR  ALNSSLEC-
   1bpm#143  VSAKL  PSVEVDP  ---EGAVLG
 1cjm_A#190  VLYLF  VVYVARN  ---QEWWEL
   1cy4_A#5  KALVI  DYVVKSS  SELKQLAEK
 1d4g_A#382  VGVHF  VPAINVN  TGVHNLYKM
  1dos_A#11  PGVIT  VPVILHT  SGAHHVHQM
```

The signal for amino acid (substitution group) *j* in position *i*:

$$s_{ij} = \frac{\hat{P}_{ij} - P_j}{\sigma_j}$$

$\hat{P}_{ij}$ - observed frequency probability
P_j - a priori probability
σ_j - standard deviation

The signal vector for segment *k*:

$$\vec{s}_k = \left\langle \underbrace{s_{11}, s_{12}, \ldots}_{\text{Position 1}}, \underbrace{s_{21}, s_{22}, \ldots}_{\text{Position 2}} \right\rangle_k \qquad k \in \{1,2,\ldots\}$$

Equivalent to a position specific scoring matrix (PSSM)

Signal matching

```
                DESCRIPTOR GROUP
  1ay7_B#81  ITIIL  LVLEW--  SVLQVFREA
 1b9y_C#114  GFVYE  VKFCKIR  ALNSSLEC-
   1bpm#143  VSAKL  PSVEVDP  ---EGAVLG
 1cjm_A#190  VLYLF  VVYVARN  ---QEWWEL
   1cy4_A#5  KALVI  DYVVKSS  SELKQLAEK
 1d4g_A#382  VGVHF  VPAINVN  TGVHNLYKM
  1dos_A#11  PGVIT  VPVILHT  SGAHHVHQM
         SS  EEEEE  EEEEEEE  HHHHHHHHH
```

$\vec{S}_1$ $\vec{S}_2$ $\vec{S}_3$

```
KKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGCDITIILS
KHCTISGRAVHSLDELYDEIARQLPLPDYFGRNLDALWDVLSTDIEGPVELIWEDSEHSKRSMGKDYERVVALLKDLTEEREDFRIV
    IIGSKIYTEQDFHNQISKIFSIQDYYGNNLDALWDLLSTNVERPITLVWKDAMFSKNQLENIFIEIVNVLERVKKQDED
          QSKQEVLETIATSFLFPKHFGKNYDALYDCLTDLVQFVIVLE--QLPVAQKFDKEGRETLLDVFREA
CEEEEECCCCCHHHHHHHHHHHCCCHHHCCHHHHHHHHHHCCCCCEEEEECCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCEEEEEC
```

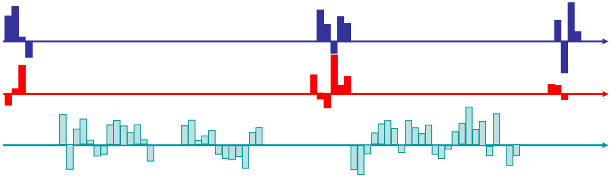
- Target sequence
- Aligned sequences (PSI-BLAST)
- Predicted secondary structure (PSIPRED)
  - H=helix
  - E=extended beta strand
  - C=coil

Signal matching

```
                DESCRIPTOR GROUP
  1ay7_B#81  ITIIL  LVLEW--  SVLQVFREA
 1b9y_C#114  GFVYE  VKFCKIR  ALNSSLEC-
   1bpm#143  VSAKL  PSVEVDP  ---EGAVLG
 1cjm_A#190  VLYLF  VVYVARN  ---QEWWEL
   1cy4_A#5  KALVI  DYVVKSS  SELKQLAEK
 1d4g_A#382  VGVHF  VPAINVN  TGVHNLYKM
  1dos_A#11  PGVIT  VPVILHT  SGAHHVHQM
         SS  EEEEE  EEEEEEE  HHHHHHHHH
```

$\vec{S}_1$ $\vec{S}_2$ $\vec{S}_3$

```
KKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGCDITIILS
KHCTISGRAVHSLDELYDEIARQLPLPDYFGRNLDALWDVLSTDIEGPVELIWEDSEHSKRSMGKDYERVVALLKDLTEEREDFRIV
    IIGSKIYTEQDFHNQISKIFSIQDYYGNNLDALWDLLSTNVERPITLVWKDAMFSKNQLENIFIEIVNVLERVKKQDED
          QSKQEVLETIATSFLFPKHFGKNYDALYDCLTDLVQFVIVLE--QLPVAQKFDKEGRETLLDVFREA
CEEEEECCCCCHHHHHHHHHHHCCCHHHCCHHHHHHHHHHCCCCCEEEEECCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCEEEEEC
```

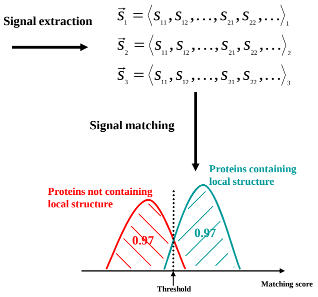


Matching score between a group and a target sequence in general

$$m = \sum_k \max_p \left\{ \vec{s}_k \cdot \vec{t}_{\subset p} \right\}$$

Method

```
          DESCRIPTOR GROUP
 1ay7_B#81  ITIIL  LVLEW--  SVLQVFREA
1b9y_C#114  GFVYE  VKFCKIR  ALNSSLEC-
  1bpm#143  VSAKL  PSVEVDP  ---EGAVLG
1cjm_A#190  VLYLF  VVYVARN  ---QEWWEL
  1cy4_A#5  KALVI  DYVVKSS  SELKQLAEK
1d4g_A#382  VGVHF  VPAINVN  TGVHNLYKM
 1dos_A#11  PGVIT  VPVILHT  SGAHHVHQM
1e5d_A#256  KVVIF  TVKLMWC  SQIMSEISD
1ecr_A#258  PTPLI  LPGVLCY  ---LRHFRH
 1ekq_B#21  LVHSI  SPVMAYA  EEVADMAKI
1eok_A#269  GGMMI  MVFGAYA  ANDVEVAKW
```


Signal extraction
Signal matching
Proteins containing local structure
Proteins not containing local structure
0.97
0.97
Threshold
Matching score

$$\vec{S}_1 = \langle S_{11}, S_{12}, \ldots, S_{21}, S_{22}, \ldots \rangle_1$$

$$\vec{S}_2 = \langle S_{11}, S_{12}, \ldots, S_{21}, S_{22}, \ldots \rangle_2$$

$$\vec{S}_3 = \langle S_{11}, S_{12}, \ldots, S_{21}, S_{22}, \ldots \rangle_3$$

Fold recognition

- ➢ Match each group (local protein structure) to the target sequence
- ➢ Assign groups with a score higher than the threshold
- ➢ Rank folds according to P-values

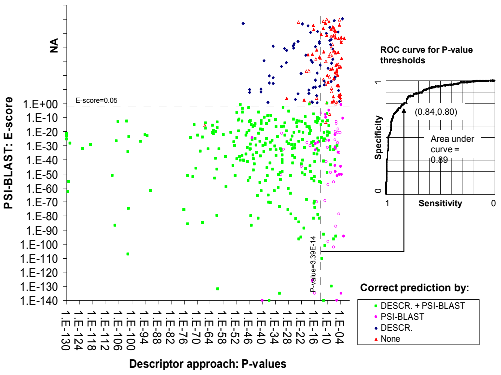
Domain: **1e9ra_**

SCOP fold: P-loop containing nucleoside triphosphate hydrolases (**c37**)

Fold	Assignment	P-value
1. c37	(41/113)	5.324e-36
2. d159	(3/7)	0.0008695
3. c66	(5/40)	0.0066181
4. c7	(2/8)	0.0226421
5. c2	(112/186)	0.0240892
6. b82	(3/26)	0.0425178
…		
30. d153	(1/77)	0.9088401

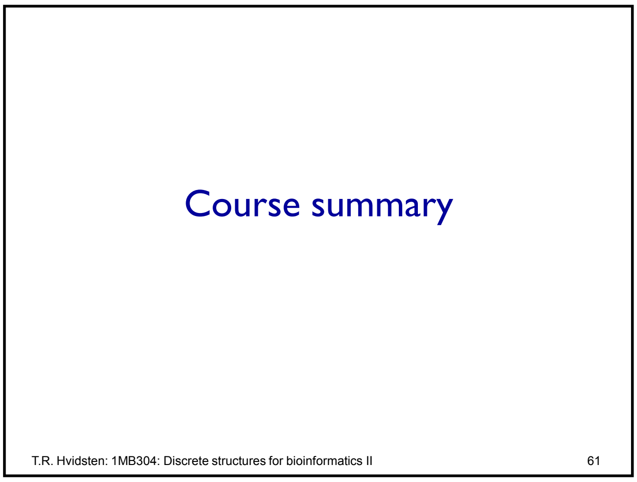
Results

479 domains with less than 40% sequence identity to the training set

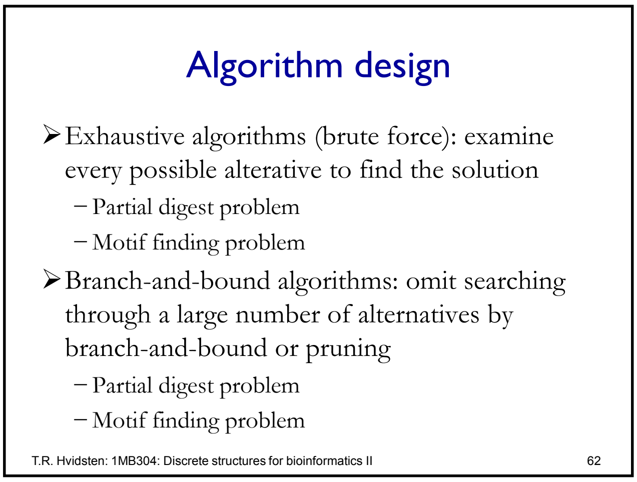


Project

- ➢ Modeling each descriptor group with an HMM rather than a set of PSSMs (profiles)
- ➢ Use HMMs to assign local substructures to new proteins
- ➢ Fold recognition

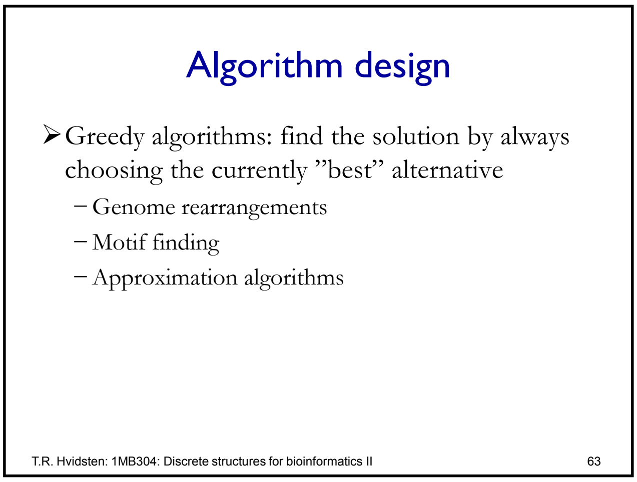


Course summary



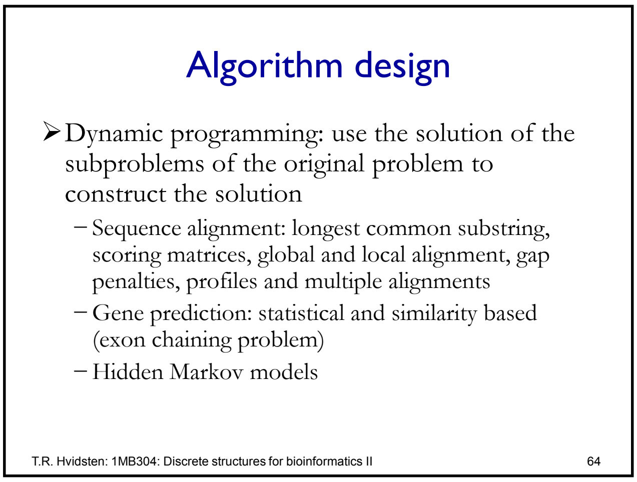
Algorithm design

- ➢ Exhaustive algorithms (brute force): examine every possible alterative to find the solution
  - – Partial digest problem
  - – Motif finding problem
- ➢ Branch-and-bound algorithms: omit searching through a large number of alternatives by branch-and-bound or pruning
  - – Partial digest problem
  - – Motif finding problem



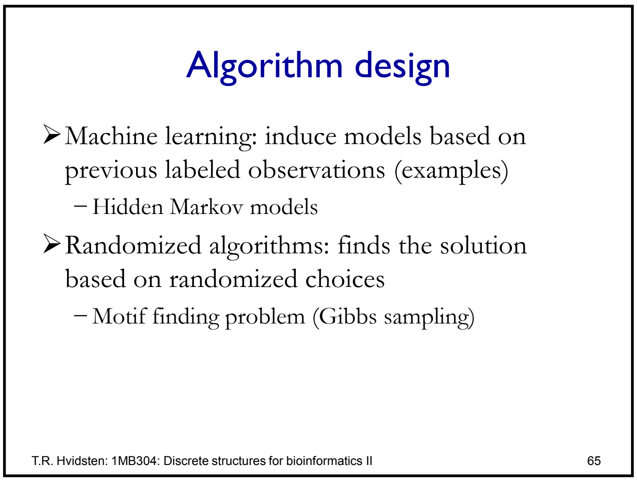
Algorithm design

- ➢ Greedy algorithms: find the solution by always choosing the currently ”best” alternative
  - – Genome rearrangements
  - – Motif finding
  - – Approximation algorithms



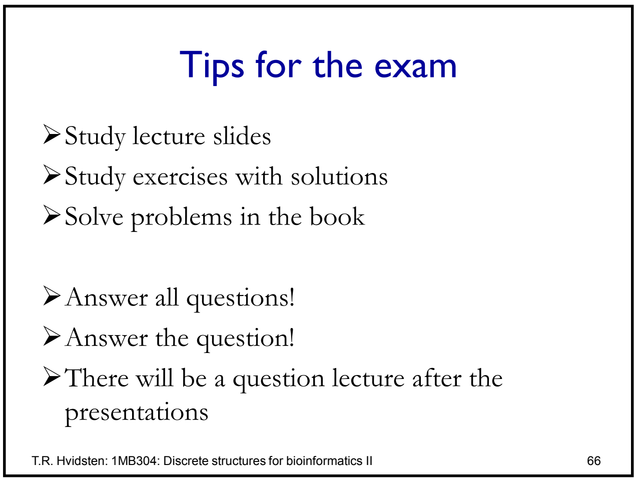
Algorithm design

- ➢ Dynamic programming: use the solution of the subproblems of the original problem to construct the solution
  - – Sequence alignment: longest common substring, scoring matrices, global and local alignment, gap penalties, profiles and multiple alignments
  - – Gene prediction: statistical and similarity based (exon chaining problem)
  - – Hidden Markov models



Algorithm design

- Machine learning: induce models based on previous labeled observations (examples)
  - Hidden Markov models
- Randomized algorithms: finds the solution based on randomized choices
  - Motif finding problem (Gibbs sampling)



Tips for the exam

- Study lecture slides
- Study exercises with solutions
- Solve problems in the book

- Answer all questions!
- Answer the question!
- There will be a question lecture after the presentations