Bioinformatics

Structural and global properties

Martin Saturka

http://www.bioplexity.org/lectures/

EBI version 0.52

Structures

Sequence based secondary structures, domains and folding

- computational structure predictions
- experimental data based explorations

Structure explorations

- dynamic programming
  - secondary structure prediction
  - RNA folds, CM and SCFG
- structural biology
  - experiments, comparisons
  - combinatorial chemistry, docking
- feature characteristics
  - hydrophobic packing
  - global properties

Structure hierarchy

from sequences to 3D structures

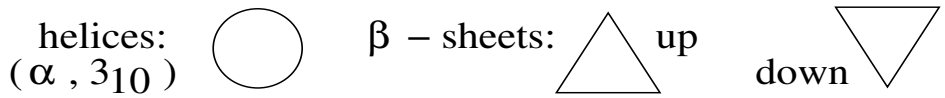
primary, secondary, tertiary, quaternary structures

- primary - linear sequences of monomers
  - modifications: crosslinking, cleavage, ligation
- secondary: local structural motifs
  - regular simple sturctures vs. random coils
- tertiary: whole single molecule structures
  - folds, complex, dependency on environments
- quaternary: molecular complexes
  - enzymatic complexes, cytoskeleton, capsids

Secondary structure

systematics of local structures of proteins and nucleic acids

- proteins
  - Ramachandran plot of dihedral angles
  - α-helices, β-sheets, coils, others

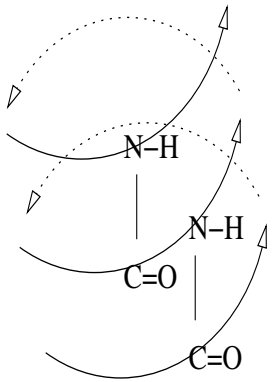


- RNAs
  - base pairing: hairpins - stem / loop
  - pseudoknots, kissing structures

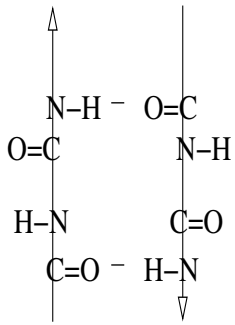
- DNAs
  - relatively rigid double helix
  - G and C quadruplex structures

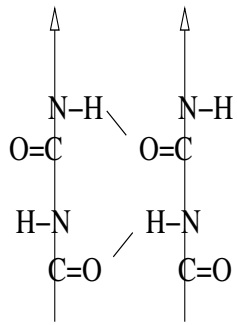
Protein local structures

the basic protein building blocks



α – helix





β – sheets

- α-helix ($\psi(i) + \varphi(i+1) \doteq -105°$)
  - right-handed, 3.6 residues per turn
  - approximately each fourth residuum to the same direction
- β-sheets
  - parallel (-120°, 115°) and anti-parallel (-140°, 135°) cases
  - alternating residuum directions, with respect to the plane

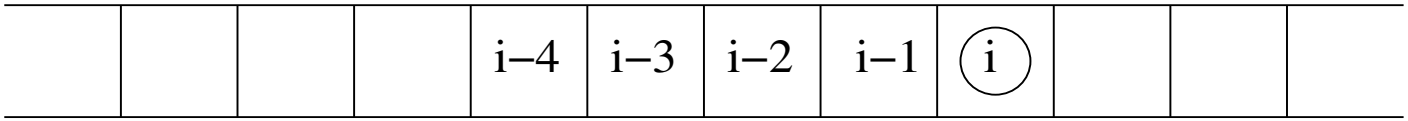
Protein local predictions

mediocre results, methods based on dynamic programming

- specifics to consider
  - start at the N' ends - first folded
  - proneness to helices, beta sheets, structure breakers
  - assumed residua by multiple sequence alignment
  - feature - e.g. hydrophobic character alternations

- altered HMM algorithm
  - *n*-th order Markov model
  - inner states: helices, sheets, turns, coils
  - consider the shortest structure lengths
    - either to look the shortest lengths backward or multiple inner states - for (short) structure lengths

Prediction example



Inner states for :
- H – helices
- E – sheets
- T – turns

H_1 H_2 H_3 ... H_n

E_1 E_2 E_3 ... E_n

T_1 T_2 T_3 T_4 T_5

α-helix, min 4 residues

propensities: `MALEK`

β-sheet, min 2 (5 $\Uparrow$) residues

propensities: `YFWTVI`

turns, 3-5 residues

propensities: `GP`

A Q G L A E

OK: H_1 H_2 H_3 H_4 H_n H_n

NO: H_1 H_2 T_1 H_1 H_2 H_3

Hierarchical HMMs

secondary and super-secondary structures

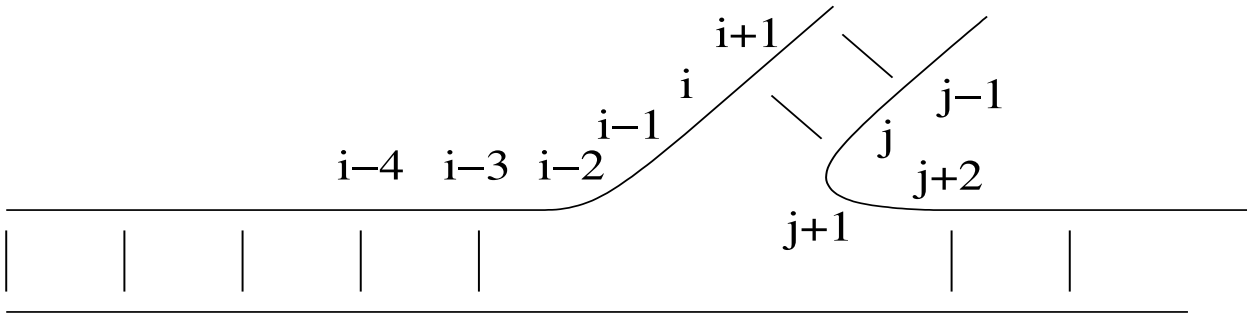
- make secondary structure prediction
  - input (outer) string: primary structure
  - inner states sequence: secondary structure
- run subsequent higher order HMM prediction
  - input (outer) string: secondary structure
  - inner states sequence: complex motifs

```
...AGPGAQGLAE...   →   ...HTTTHHHHHH...
...HTTTHHHHHH...   →        helix-turn-helix
```

nucleotides pairing - strong long range interactions

standard and non-standard ribonucleotide pairing

- HMMs: without long range interactions, not sufficient
- covariance versions of locality based algorithms
  - covariance Gibbs sampling - pairs of trials
  - covariance extension of HMMs - pairwise probabilities



Automata and grammars

- Chomsky-Schützenberger hierarchy
  - regular languages
  - context-free languages
    - non-terminal symbols (with the start symbol)
    - terminal symbols, the outer alphabet
    - rewrite rules
  - context-sensitive languages
  - recursive languages
- stochastic models
  - finite automata / regular grammars $\longrightarrow$ HMMs
  - pushdown automata / context-free grammars $\longrightarrow$ CMs

nonterminals: s, a_1, a_2 terminals: A, C, G, U

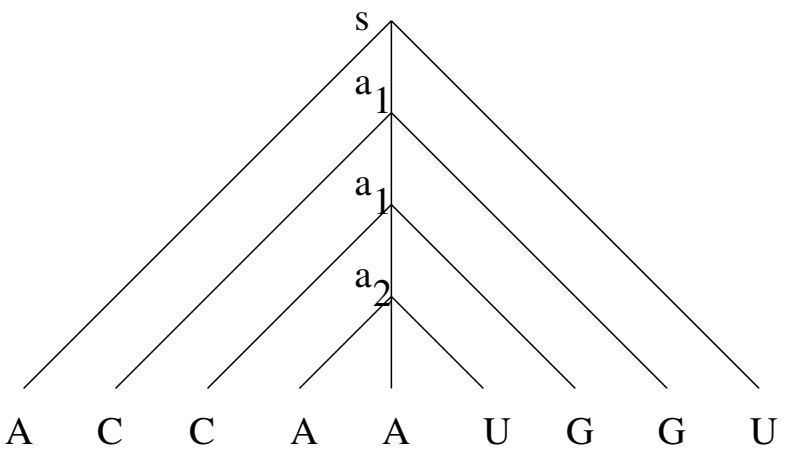
$$
\begin{aligned}
s &\longrightarrow A\ a_1\ U \\
a_1 &\longrightarrow C\ a_1\ G \\
a_1 &\longrightarrow C\ a_2\ G \\
a_2 &\longrightarrow A\ A\ U
\end{aligned}
$$

ACC...CCAAUGG...GGU

Covariance model

SCFG (PCFG): stochastic (probabilistic) context-free grammars

- sets of: terminals, nonterminal, probabilistic rules
- probability of rule sums for each nonterminal is unite
  - rewrite rules play roles of both inner transitions and symbol outputs of HMMs



60%: $a_1 \longrightarrow C\ a_1\ G$

10%: $a_1 \longrightarrow A\ a_1\ A\ a_1\ A$

30%: $a_1 \longrightarrow C\ a_2\ G$

CM algorithms

time complexity of n^3 for lengths of parsed sequences

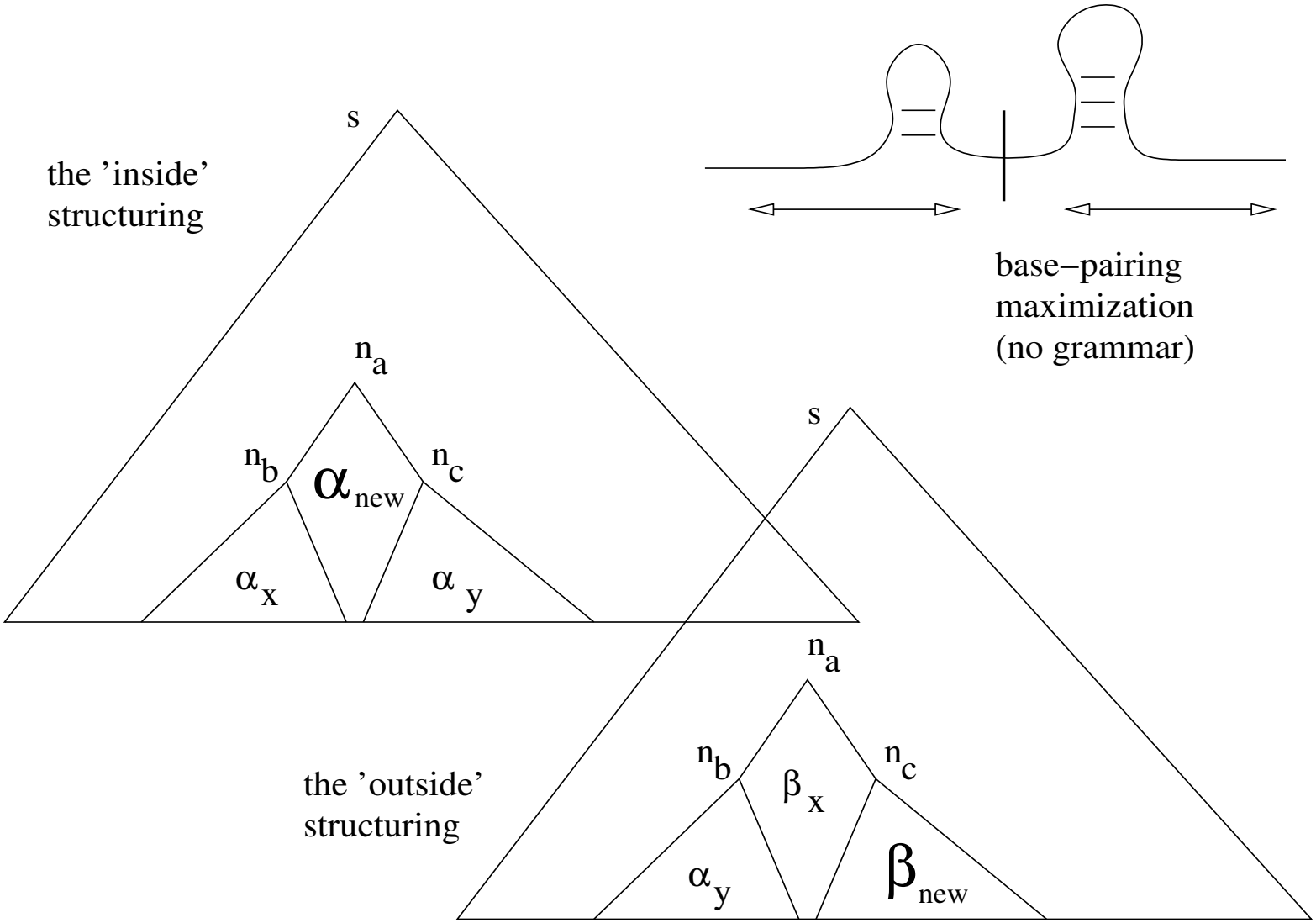
- weighted CYK algorithm for the most probable production
  - analogy of the Viterbi algorithm of HMMs
- inside / outside algorithm for SCFG adjusting
  - analogy of the forward / backward algorithm of HMMs

- the Inside and weighted CYK algorithms
  - difference: 'inside' makes sums, 'CYK' takes maxima
  - iterative substring parse generation (for the CYK)
    - first, finding the best parses for short subsequences
    - for larger subsequences, make the best separation onto the most paired / the best parsed subsequences
    - possibility to do separations at just single points

Inside algorithm

- normalizing grammar rules for just binary tree parses
  - $n_a \rightarrow n_b n_c$ - for inner state changes
  - $n_a \rightarrow T_r$ - for outer symbol outputs
- square matrices indexed by sequence positions
  - each matrix for subsequences from a single non-terminal
  - just one matrix for a non-grammar / best pairing search
- first, filled with zeros - initial parse weight sums
  - diagonals with probabilities of respective symbol outputs
- iterative matrices filling out of the main diagonals
  - computation for each matrix of a nonterminal symbol
  - for every subsequence make its two-pieces separation
  - compute parse weights for the pieces of the subsequence for generation form any pair of nonterminal symbols
  - multiply with the probability of the nonterminal pair rule
  - the end is for the whole sequence from the start symbol

Algorithm structuring



Inside / Outside

- the Inside algorithm: $\alpha(i, j, n_a)$
  - probability sums of all parse trees of subsequence (i to j positions) generated from the n_a nonterminal
- the Outside algorithm: $\beta(i, j, n_a)$
  - probability sums of all parse trees without counting the probabilities of the (i to j positions) subsequence generation from the n_a nonterminal

$$
\begin{aligned}
\alpha(i, i, n_a) &= \Pr(n_a \rightarrow o(i)) \\
\alpha(i, j, n_a) &= \sum_{n_b} \sum_{n_c} \sum_{k=i}^{j-1} \alpha(i, k, n_b) \cdot \alpha(k+1, j, n_c) \cdot \Pr(n_a \rightarrow n_b\, n_c)
\end{aligned}
$$

$$
\begin{aligned}
\beta(1, |o|, n_s) &= 1 \text{ for the start non-terminal} \\
\beta(1, |o|, n_z) &= 1 \text{ for a non-start non-terminal} \\
\beta(i, i, n_a) &= \sum_{n_b} \sum_{n_c} \sum_{k=1}^{i-1} \alpha(k, i-1, n_b) \cdot \beta(k, j, n_c) \cdot \Pr(n_c \rightarrow n_a\, n_b) + \\
&\quad \sum_{n_b} \sum_{n_c} \sum_{k=j+1}^{|o|} \alpha(j+1, k, n_b) \cdot \beta(i, k, n_c) \cdot \Pr(n_c \rightarrow n_b\, n_a)
\end{aligned}
$$

CM profiling

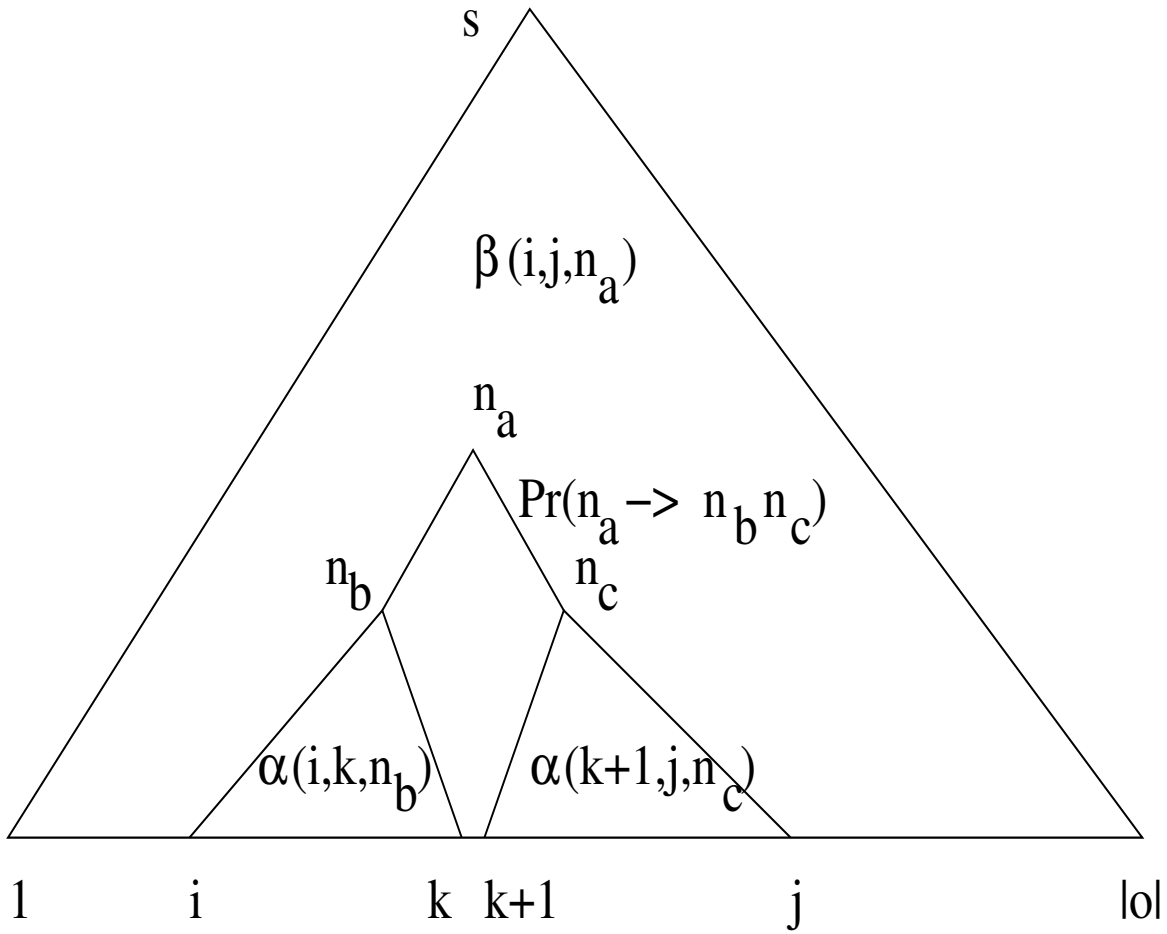
- the covariance model can work without a grammar
  - just with maximizing base pairing - poor results
- having a grammar, we need to adjust the probabilites
  - analogically to HMM profiling

- parameter reestimation by expected times of a rule usage
  - divided by all the rules usage from the non-terminal
- new output probabilities
  - new $\Pr(n_a \rightarrow T_r) = c(n_a \rightarrow T_r)/c(n_a)$
  - count of n_a used to generate the terminal T_r
    $c(n_a \rightarrow T_r) = \sum_{i,o(i)=T_r} \beta(i,i,n_a) \cdot \Pr(n_a \rightarrow T_r)$
  - count of n_a used to generate anything
    $c(n_a) = \sum_i \sum_j \beta(i,j,n_a) \cdot \alpha(i,j,n_a)$

Profiling counts

- new $\Pr(n_a \rightarrow n_b\, n_c) = c(n_a \rightarrow n_b\, n_c)/c(n_a)$

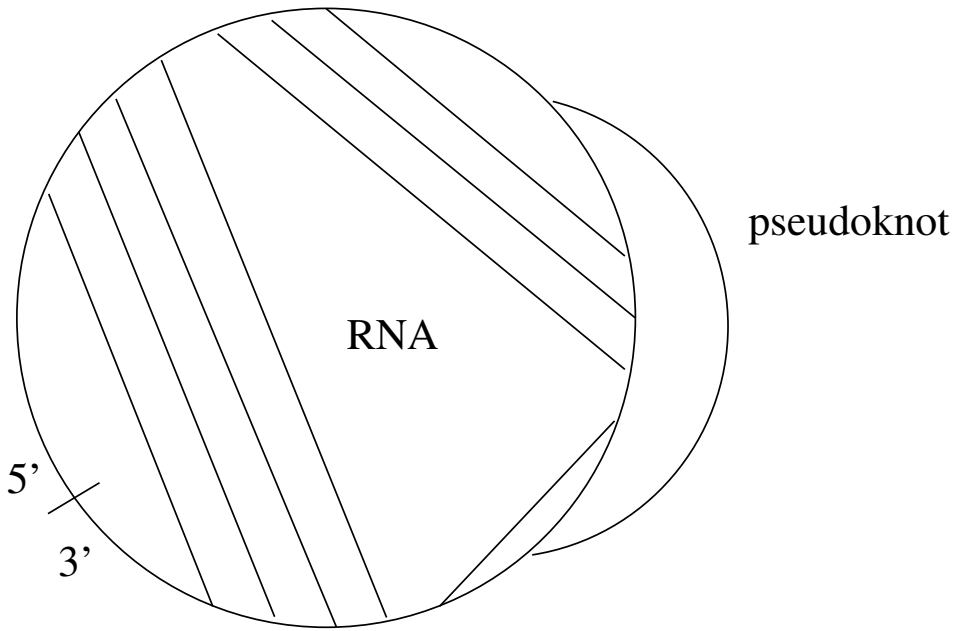
$c(n_a \rightarrow n_b\, n_c)$ the count of n_a used to generate a non-terminal pair $n_b\, n_c$ is $\sum_{i=1}^{|o|-1} \sum_{j=i+1}^{|o|} \sum_{k=i}^{j-1} \beta(i,j,n_a) \cdot \Pr(n_a \rightarrow n_b\, n_c) \cdot \alpha(i,k,n_b) \cdot \alpha(k+1,j,n_c)$



CM obstacles

- high time complexity
  - not suitable for large RNA molecules

- pseudoknots
  - usually low depth subtree separations



Physical approaches

energy minimization

- while symbolic base-pairing approaches popular, physics agnostic methods suffer from the ignorance
  - many dispersed short base pairing unfavourable
  - different base-pairs of different strengths
  - minor bases common in RNA molecules

- complex RNA structures
  - complex molecular modeling and stochastic grammars
  - alignment based structure prediction
    - many RNA molecules with known folds

Experimental technics

structural biology

necessary source of solid molecular structure data

- standard technics
  - crystallography
    - inner cores generally correct
    - reduced possibilities for surfaces and domain flipping
  - NMR spectroscopy
    - less accurate than X-ray diffraction
    - measurements in more natural environments
  - IR, Raman spectroscopies
    - simpler, for vibrations of specific parts
  - EM, AFM
    - for structures of greater molecular complexes
  - other methods
    - many kinds of spectroscopy and microscopy, ultracentrifugation, chromatography, etc.

Structure refinement

- frequent usage
  - experimental data adjustment
  - exploring small alterations
  - short macromolecular dynamics
  - states of small molecules
- molecular mechanics
  - statics, energy minimizations
  - standard hill-climbing methods
- molecular dynamics
  - intensive computer simulations
  - amount of solvent, long range interactions
- stochastic dynamics
  - Langevine dynamics - extra random forces
  - Monte Carlo - probabilities, not forces

Force fields

biomacromolecules: classical forces approximation

quantum potentials for ligands, limited areas

- environment approximation
  - charge shielding, hydrophobic interaction, entropy
- empirical potentials
  - bonds, bond angles, dihedral angles
  - electrostatic, van der Waals forces
  - implicit solvatation

Structural alignment

structure comparison and prediction

- comparing similar structures
  - minimizing root mean square of distances
  - distance matrices for chosen atoms
- sequence to structure alignment
  - prediction of structures of large protein blocks
  - popular methods with growing staructural databases
  - structural alignment onto structures of similar sequences
- protein threading
  - threading 1D sequences onto 3D structures
  - usable technics with large structural databases available
    - chance of a strucutre with a domain of a similar sequence

Enzyme activity

proteins been most studied, RNAs as the current 'big thing'

- several basic facts
  - active sites formed by sequence-distant residua
  - induced fit action - structure adjustment on substrates
  - enzyme activity modulation by cofactors and coenzymes
  - structure change as allosteric regulation of many enzymes
  - in evolution, function interchange as receptors
- protein flexibility
  - enabling huge amount of protein functions
  - path to intentionally regulate enzyme functions
  - path to escape intentional regulations

Ligand docking

- combinatorial chemistry
  - drug design - de novo, known substrate alterations
  - usage in medicinal chemistry, pharmaceutical industry
  - computational reduction of vast amount of ligands
- QSAR approaches
  - quantitative structure-activity relationship
  - rules for combinatorial ligand construction
- docking methods
  - generation of possible ligand conformations
  - initial ligand positions and orientations
  - molecular mechanics to minimize interaction energy
  - too tight bindings lack entropy contributions

Protein folding

- random vs. natural sequences
  - random polypeptides do not form folded structures
  - proteins with folded and denaturated forms
- folding paths
  - Levinthal paradox - too large amount of degrees of freedom thus sampling just a minor fractions of them possible
  - funnels of folding paths, directing to the right conformations
    - many proteins need chaperons for the right folds
    - dual forms of prions, probably of many other (innocent) proteins, hidden by cellular degradation pathways as well

Fold predictions

threading - global structure predictions

- inverse approaches more feasible than direct predictions
- threading of altered structures onto the original folds
  - generated databases of such threaded sequences

- search threading databases for similar fragments
- arrangement of the subsequences onto the structures
- scoring with coarse-grained pseudo-energy functions
  - better with experimental (e.g. NMR) distnace constraints

Global view

properties along the whole sequences

- hydrophobic character
- regulatory - active sites connection
- position vs. frequency views
- auto-correlation, repetitions

- structure recognition by hydrophobicity distribution
  - membrane proteins with specific characteristics

Sequence profiles

- float-point sequences
  - hydrophobic values, charge values
  - structure prediction by profile similarity
- cores vs. surfaces
  - hydrophobic cores as structure identification
  - convex and alpha hulls surfaces for protein docking
- wavelet analysis
  - localizing at both position and frequency spaces
  - used for protein core predictions on hydrophobic scales
    - the reliability claimed similar to that of standard secondary structure predictions

Qualitative modeling

qualitative topology and flexibility

- discrete structure modeling
  - hydrophobic packing
  - frustration minimization
- qualitative dynamics
  - coarse-grained domain vibrations
  - normal modes of the domains

Other biomacromolecules

saccharides - the next 'big thing'

- lipid membranes
  - separation of the inner vs. the outer
  - surfaces with protein and carbohydrate markers
- carbohydrates
  - major roles in immune system, cell recognition
  - common glycosylation of lipids and proteins

Items to remember

Nota bene:

molecular structure hierarchy

- Secondary structure predictions
  - proteins
  - RNAs
- Higher order structures
  - alignments
  - global methods