Structure of proteins
Modeling and drug design

Marcus Elstner and Tomáš Kubař

July 28, 2017

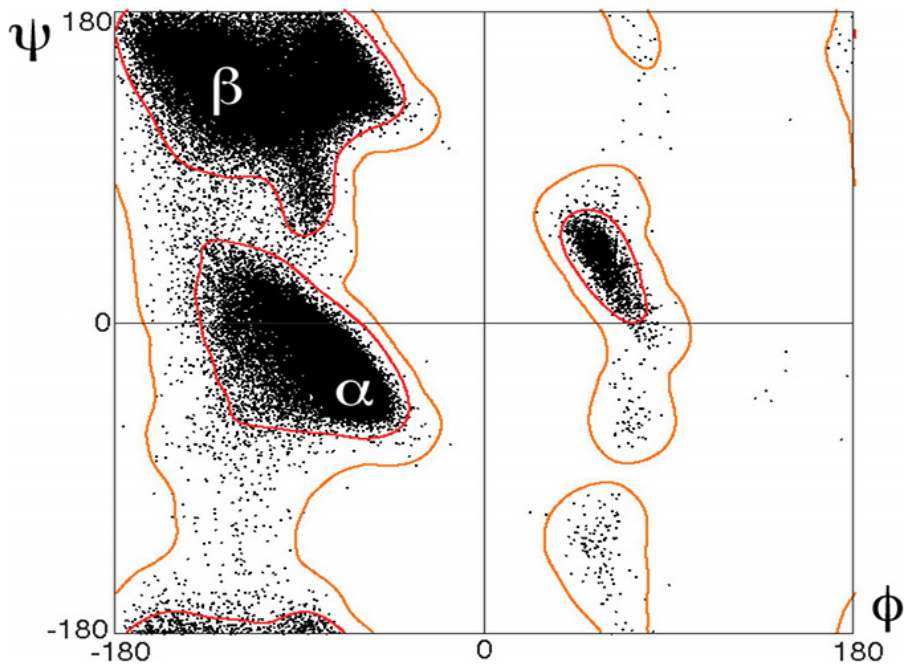
Structure of proteins

Basic principles of protein structure

- very complex, yet some structural patterns occur often
- secondary structure
  – α-helix, β-strand, rare helices, several kinds of loops / turns
- role of hydrogen bonds
- tertiary structure – orientation of 2° structures (β-barrel)
- quaternary structure – organization of individual subunits
- native, active state of a multi-subunit protein

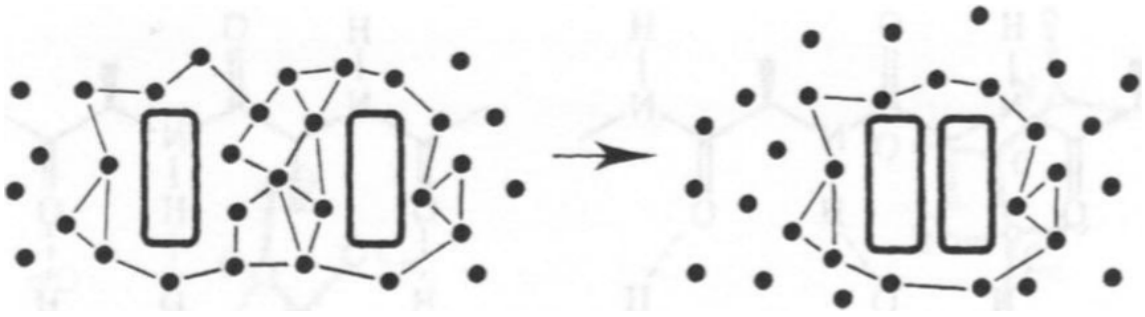
Structure of a polypeptide chain

- characterized by the dihedral angles along the backbone
- planar configuration on the amide bond
- two dihedral angles per AA – φ (N–C$^{\alpha}$) and ψ (C$^{\alpha}$–C)
- Ramachandran plot (1963); any amino acids lying outside of the common regions in RP would be paid special attention



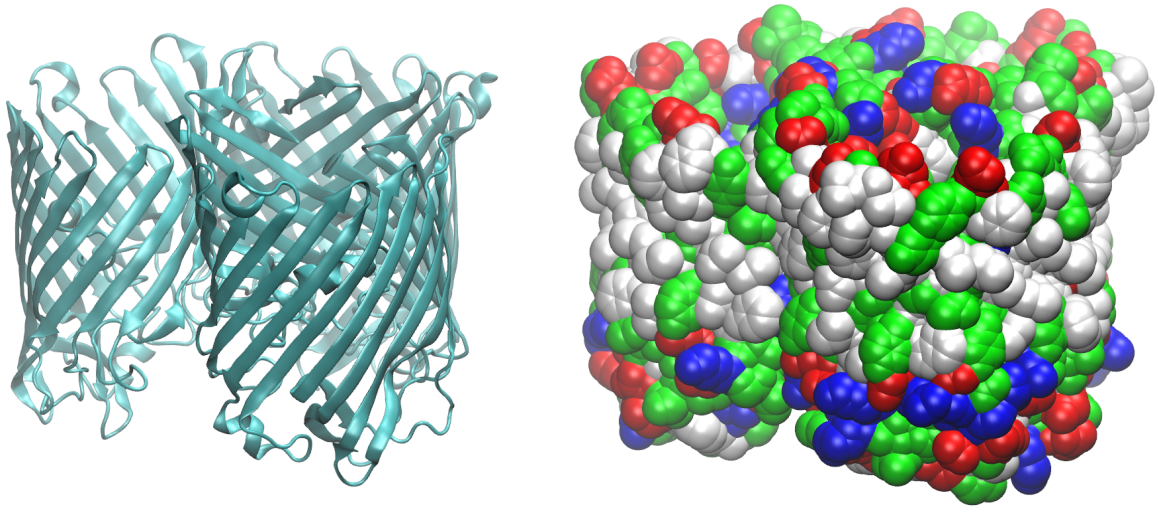
Soluble / globular proteins

- surface – polar and charged amino acids,
- non-polar AAs (Trp Phe Leu Ile Val) cumulate in the interior
- hydrophobic effect – crucial for the stability of proteins
- folding – free surface of non-polar AA side chains decreases
  - H_2O molecules are released from the 'cage' to bulk water
  - increase of entropy believed to dominate ΔG of creation of the native structure



Transmembrane proteins

- non-polar AA side chains
  - on the surface in the membrane-spanning region
  - match the hydrophobic character of the environment in the interior of the lipid membrane
- charged and polar AAs – exposed to the aqueous solution



outer-membrane ion transporter protein OmpF from *E. coli*, PDB ID 4D5U

Transmembrane proteins

- resolution of structure – difficult in general
- X-ray scattering approach
  - issues with crystallization of such proteins
- improved cryo-electron microscopy
  - certain advantages over X-ray scattering
  - under continuing development, as of 2015
  - involves huge amounts of computation

Review

CellPress

How cryo-EM is revolutionizing structural biology

Xiao-chen Bai, Greg McMullan, and Sjors H.W Scheres

MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, UK

Trends in Biochemical Sciences, January 2015, Vol. 40, No. 1 49

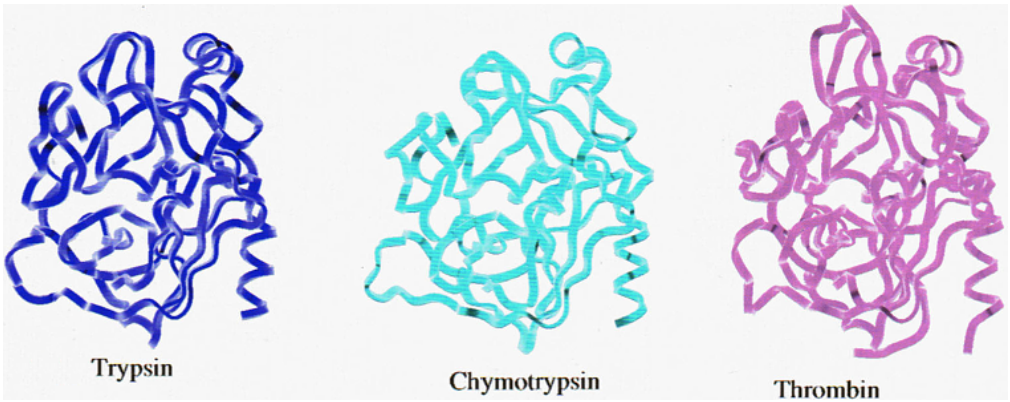
Comparative/homology modeling

- method to obtain a model of protein structure.
- to build 3D structure, based on comparison of the sequence to that of certain other protein(s)
- homology = structural similarity (+ evolutionary origin)
- protein structures are more conserved than protein sequences
- often: similar AA sequence → nearly identical 3D structure

```
Trypsin        SQWVVSAAHC ..... YKSGIQVRLG EDNINVVEGN E.QFISASKS
Chymotrypsin   EDWVVTAAHC ..... GVTTSDVVVA GEFDQGLETE DTQVLKIGKV
Thrombin       DRWVLTAAHC LLYPPWDKNF TVDDLLVRIG KHSRTRYERK VEKISMLDKI

Trypsin        IVHPSYN.SN TLNNDIMLIK LKSAASLNSR VASISLP... TSCA..SAGT
Chymotrypsin   FKNPKFS.IL TVRNDITLLK LATPAQFSET VSAVCLP... SADEDFPAGM
Thrombin       YIHPRYNWKE NLDRDIALLK LKRPIELSDY IHPVCLPDKQ TAAKLLHAGF

Trypsin        QCLISGWGN. ....TKSSGT SYPDVLKCLK APILSDSSCK SAYPGQITSN
Chymotrypsin   LCATTGWGK. ....TKYNAL KTPDKLQQAT LPIVSNTDCR KYWGSRVTDV
Thrombin       KGRVTGWGNR RETWTTSVAE VQPSVLQVVN LPLVERPVCK ASTRIRITDN

Trypsin        MFCAGYLEGG ...KDSCQGD SGGPVV..CS GK....LQGI VSWGSGCAQK
Chymotrypsin   MICAG..ASG ...VSSCMGD SGGPLV..CQ KNGAWTLAGI VSWGSSTCST
Thrombin       MFCAGYKPGE GKRGDACEGD SGGPFVMKSP YNNRWYQMGI VSWGEGCDRD
```

Comparative/homology modeling

- method to obtain a model of protein structure.
- to build 3D structure, based on comparison of the sequence to that of certain other protein(s)
- homology = structural similarity (+ evolutionary origin)
- protein structures are more conserved than protein sequences
- often: similar AA sequence → nearly identical 3D structure



Comparative/homology modeling

1. Identify a template – protein that we consider homologous
   (there may be more than one template)
2. Align the sequences
   – lay them next to each other to maximize the match
3. Identify which regions are structurally conserved,
   and which are probably variable regions
4. Create a model (coordinates) of the conserved region – 'core'
5. Generate the structure of the variable region(s)
   – often no regular 2° structure
6. Handle the AA side chains
7. Verify structure, and possibly refine (with MD)

Comparative/homology modeling

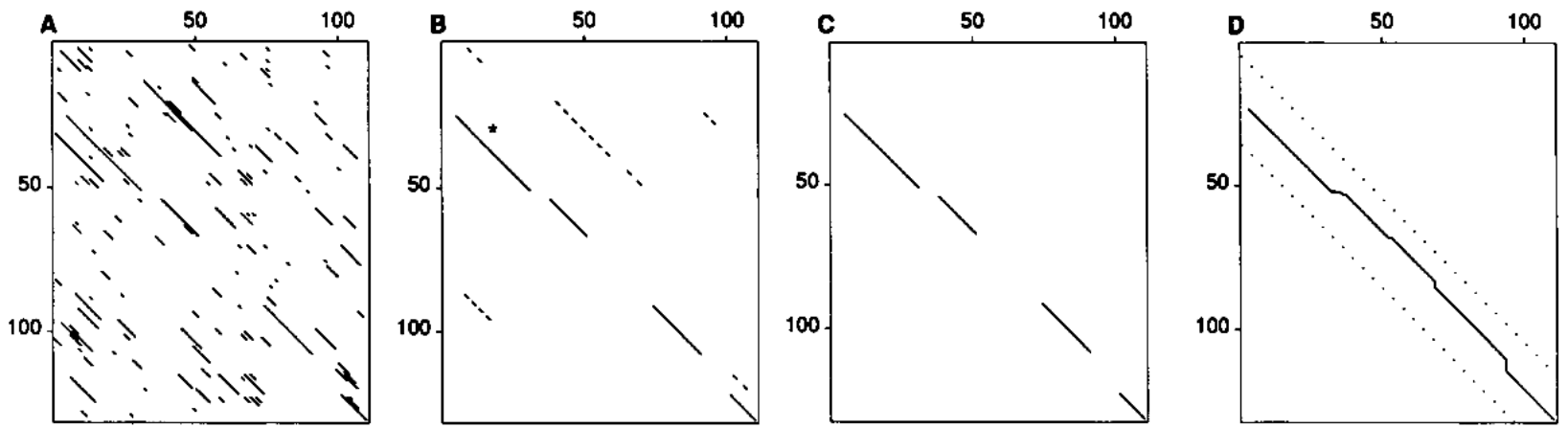
Identification of the template

- template = a protein we expect to have very similar structure
- AA sequence – the only input → comparison of the sequence with database of proteins with known 3D structure.
- usual choice – one or more proteins with sequence similarity
- also potentially useful – look for a possible function
  – provides a hint to strongly conserved fragments of sequences
  – AAs binding a cofactor or catalytic sites

Comparative/homology modeling

Alignment of the sequences
– crucial and highly non-trivial
– choice of algorithm, scoring method, application of gap penalties
– available: FASTA (quick), Smith–Waterman, BLASTP (no gaps)

1. locate regions of identity
2. scan them with a scoring matrix and save the best matches
3. optimally join initial regions to give a single alignment
4. reoptimize alignment, centered around the best scoring region



Comparative/homology modeling

Scoring of the alignment

- scoring method – gives the quality of alignment numerically
- generally: AA identical → high contribution to the score
  AAs similar (conservative) but not identical → lower score
  very different AAs aligned → unfavorable score

several possibilities to perform the scoring:

- identity – only identical AAs have favorable score.
- genetic code – score = number of nucleobase changes in DNA needed to change the AA to the other/aligned one
- chemical similarity – considering chemical properties of AAs
  example: Glu aligned with Asp scores high
- observed substitutions – based on the frequency of mutations in the alignment of sequences in protein databases

Comparative/homology modeling

Scoring functions – observed substitutions

- considered to be the best choice
- 'percentage of acceptable point mutations' (Dayhoff 1978)
  – prob. of certain mutation within evolutionary time interval
  – evol. time may be varied – different range of mutations
- scoring the alignment of 3D structures rather than sequences
  – JO matrices (Johnson & Overington 1993)
  – potentially more sensitive to similarities of 3D structures, even if the sequences are formally less similar
- no ultimate scoring approach, most suitable universally
  – selection of the scoring matrix is non-trivial
- another decision pending – global alignment (whole seq.) $\times$ local alignm. (fragment(s)) – more freedom with templates

Comparative/homology modeling

Scoring functions – observed substitutions

% probability that AA in col. *j* will have mutated to AA in row *i* by the end of the period of 250 PAM:

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp	D	5	3	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln	Q	3	5	5	6	1	10	7	3	8	2	3	5	3	1	4	3	3	1	2	2
Glu	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met	M	1	1	1	1	0	1	1	1	1	2	3	2	7	2	1	1	1	1	1	2
Phe	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	19	6	5	1	2	4
Ser	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val	V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Comparative/homology modeling

Gap penalty

- the alignment of sequences is allowed to be discontinuous
- this is penalized by an unfavorable contribution to the score
- simplest way: a constant negative contribution for each indel
- better: penalty $= u + v \cdot k$ for a gap of length k AAs
  opening penalty $u >$ extension penalty v
- even more complex: apply larger penalty
  if the gap lies within a 2° structure element
  or even within an active center of the protein

Comparative/homology modeling

Structurally conserved/variable regions (CR/VR)

- CRs – assumed the same 3D structure in the unknown protein
- VRs – will require special treatment afterwards
- more feasible if more than one template is available
- CRs – usually 2° structure elements and binding sites; these can be recognized even with only one template
- more than one template? align them first with each other $\rightarrow$ identify CRs among templates $\rightarrow$ align the unknown protein

Comparative/homology modeling

Create the 3D structural model

- generate the main chain in CRs – simply take the template(s)
- side chains – still easy for identical/similar AAs
- larger difference AAs in CRs – some systematic approach to obtain a model of side chain – e.g. rotamer libraries – take one of the most favorable conformations
- VRs – more difficult – can be still copied if sequence similar
- VRs not similar – look up the sequence among all proteins
- quite likely – no perfect match,
  large effort in application of rotamer libraries

DB of structures from C/HM – ModBase, SwissModel Repository
SwissModel via ExPASy web server, What If via EMBL servers

Evaluation and refinement of the generated structure

- large amount of knowledge of protein structure → fundamental principles established, empirical rules derived

possible criteria to check how reasonable the generated model is:

- main chain in expected regions of the Ramachandran plot
- planar peptide bonds
- side chains in accordance with prev. observation / rotamer lib.
- polar groups H-bonded to suitable partners if buried inside
- reasonable match between hydrophilic/-phobic side chains, possibly H-bonding between polar side chains and backbone
- no unfavorable atom–atom contacts (clashes)
- no empty space in the interior of the structure

programs available – Procheck, 3D-Profiler
the analysis may point at the suspicious regions of the structure

Evaluation and refinement of the generated structure

Final refinement of the structural model

- MM energy minimization, probably MD
- VRs free to move, restraints on CRs at the start;
  restraints decreased/removed during the process, gradually
- consideration of solvent – implicit/explicit, maybe PBC
  crystallographic H_2O in CRs of template may be introduced

Molecular modeling in the drug design

Drug design

One of the most exquisite applications of molecular modeling
– construct new molecules to interact in a defined way
with natural materials – proteins, nucl. acids, carbohydrates. . .

typical – 'drug design' – find a potent inhibitor of an enzyme,
not interacting harmfully with other substances in the organism

difficulties:

- the drug has to be a potent inhibitor of the given enzyme
- but it must not interact with other enzymes (possibly lethal)
- it must not decompose too early
- its metabolites must not be (too) toxic

"All things are poison and nothing is without poison;
only the dose makes a thing not a poison." Paracelsus (1493–1541)

Drug design

hard and expensive business – several hundred million € per drug

"A pharmaceutical company utilizing computational drug design
is like an organic chemist utilizing an NMR.
It won't solve all of your problems,
but you are much better off with it than without it."

(David C. Young – tables / pictures from his book follow)

"Fail good, fail early."

Drug design

TABLE 1.1 Typical Costs of Experiments

Experiment	Typical Cost per Compound ($)
Computer modeling	10
Biochemical assay	400
Cell culture assay	4,000
Rat acute toxicity	12,000
Protein crystal structure	100,000
Animal efficacy trial	300,000
Rat 2-year chronic oral toxicity	800,000
Human clinical trial	500,000,000

Moore's law: The number of transistors that can be placed inexpensively on integrated circuits doubles every $1^1/_2$ to 2 years.

Eroom's law: The number of new drugs approved per billion US$ spent on R&D has halved roughly every nine years since 1950.

Drug design

Molecular receptors – usually proteins that are important for the processes taking place in the cell

Some estimates:

- 70 % of receptors – members of 10 protein families
- 50 % of receptors – 4 families:
  GPCR, nuclear receptors, ligand-, voltage-gated ion channels
- CATH (Class, Architecture, Topology, Homology database) says: there are ca. 130 druggable domains

Drug design

Note: We will concentrate on structure-based drug design (SBDD)

Another approach – ligand-based drug design (LBDD)
– if active ligands are already known:

- any number of ligands
  – looking for similar molecules (2D or 3D)
- a few ligands
  – looking for a motif in the ligands – pharmacophore
- many ligands (20+)
  – looking for relation between structure and activity (QSAR)

SAR paradox – even very similar molecules
sometimes possess drastically different activities

Drug design

Note: We will concentrate on structure-based drug design (SBDD)

Another approach – ligand-based drug design (LBDD)

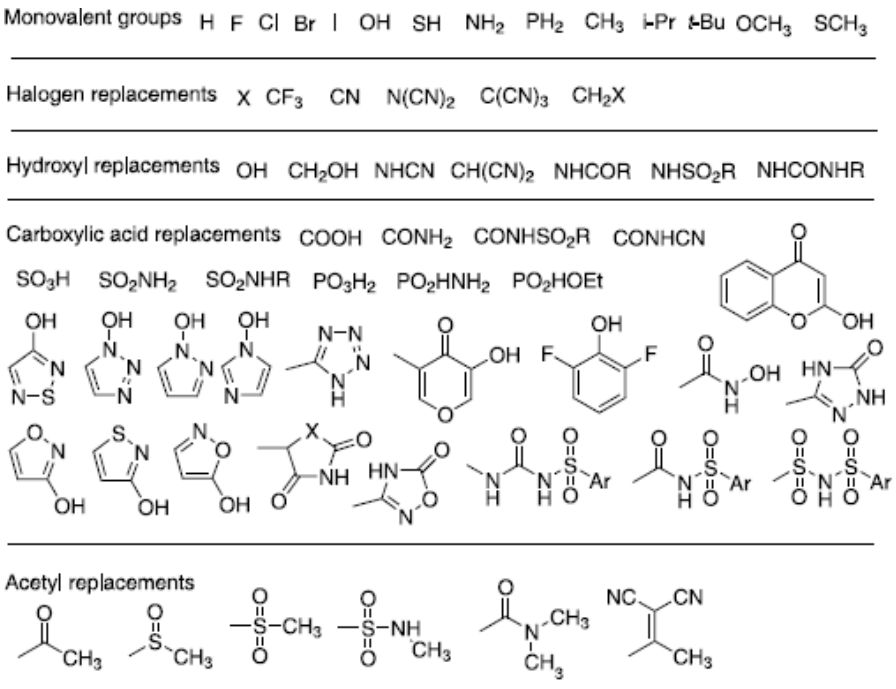
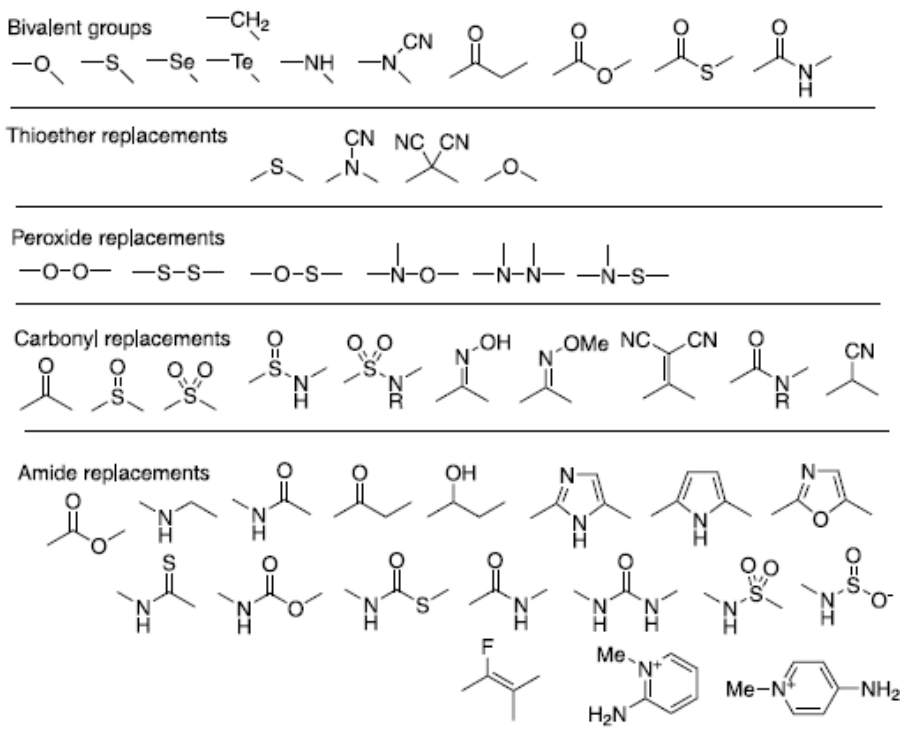


Figure 5.5 Bioisosteres.



Molecular docking

to do: find a (small) molecule (ligand, guest, key)
that would bind to a protein (receptor, host, lock)
as strongly and specifically as possible

- generate the structure of a complex of
  a known receptor (protein) and
  an up to this point unknown compound
- evaluate this structure

good news – binding site (pocket) is usually known,
often – active or allosteric place of the protein

Molecular docking

bad news:

- many degrees of freedom – trans+rot+flex of the ligand
- (relaxation of protein – may be often neglected)
- a single molecule can be docked manually
  (it helps to know the binding mode of a similar molecule)
  but not millions of molecules
- even such a straightforward approach may fail
  – even similar molecules may bind in different ways

Molecular docking

sequence of tasks to accomplish:

1. take the compounds to test from somewhere
   – database of compounds, construct from dbase of moieties. . .
2. place the molecule into binding site in the most favorable way
   – pose – orientation and conformation
3. evaluate the strength of the orientation
   accurate determination of ΔG impossible – scoring desired

Molecular docking

comment #1 on task 1:

Lipinski's rule of five / Pfizer's rule of five / RO5
for the 'druglikeness' of a compound:

there is no more than 1 violation of these conditions:

- no more than 5 hydrogen bond donors
- no more than 10 hydrogen bond acceptors
- molecular mass less than 500 daltons
- octanol-water partition coefficient $\log P \leq 5$
- (sometimes: not more than 5 rotatable bonds)

Molecular docking

comment #2 on task 1: frequently occuring moieties

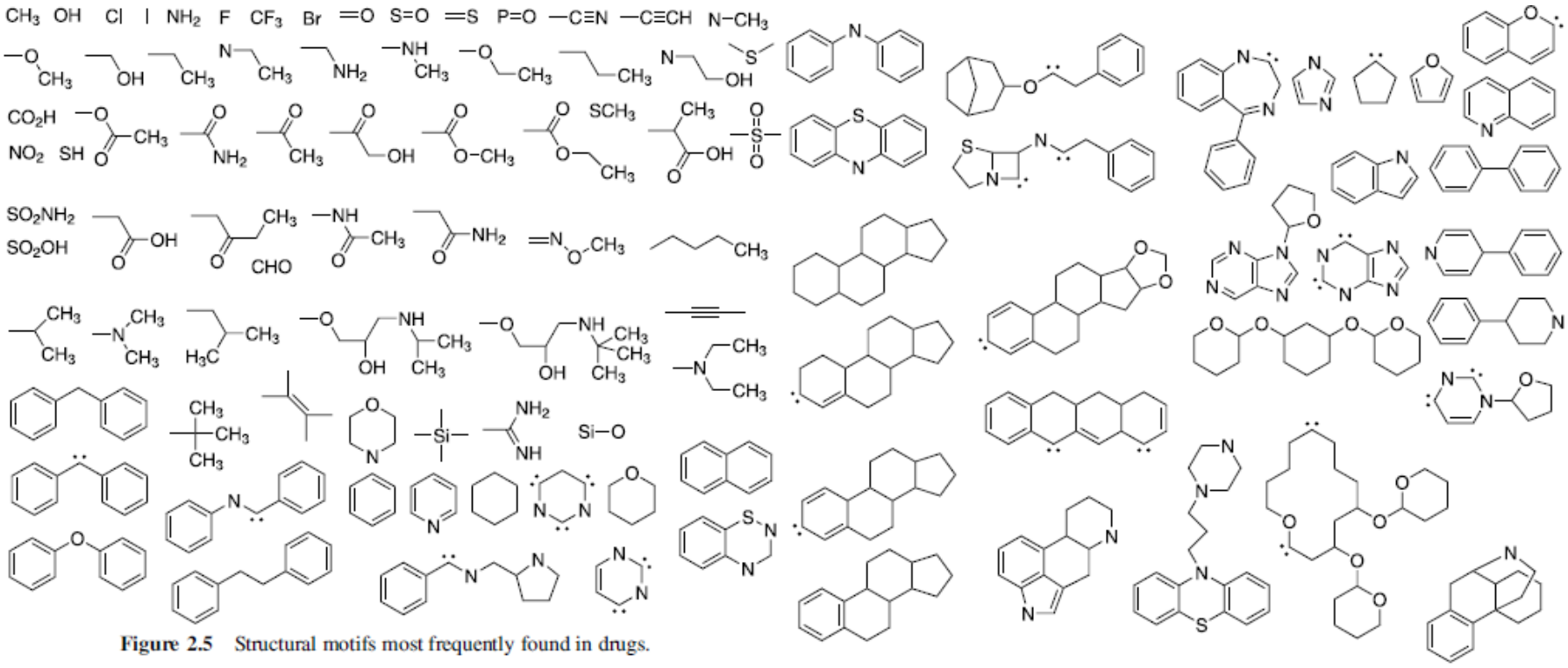


Figure 2.5 Structural motifs most frequently found in drugs.

forbidden groups: thiourea, disulfide, thiol, ester, amide, β-lactam, O-nitro, alcoxypyridinium, benzophenone, oxadiazine, fluorenone, acyl hydroquinone

Molecular docking

Various levels of approximation may be employed

The simplest approach

- process a database of molecules
- consider each of them as rigid body
- try to fit into a rigid binding pocket in the protein
- e.g. in the Dock program ('negative image' of binding pocket as a union of several spheres)

Molecular docking

Natural expansion

- consider the flexibility of the ligand in some way
- any means of exploring the config space of the molecule:
  – energy minimization, Monte Carlo,
  genetic algorithms, MD (simulated annealing)
- a simple force field

An efficient alternative

- incremental construction of the ligand
- ligand – partitioned into chemically reasonable fragments
- first fragment docked in a usual way
- other fragments are ‘grown’ consecutively
- natural way to account for the conformational flexibility
  – relative orientation of the individual fragments

Molecular docking

Problem of docking – it is all about sampling!
No way to try to do MD for every candidate molecule

- MD takes much longer than affordable – many molecules!
- MD could work probably only for quite rigid molecules
  and a binding pocket that does not constrain the ligand
  – hardly ever the case

If our goal is to dock a single, specific molecule
then a particularly thorough search with MD is possible
But if we have to dock and assess many candidate ligands
simpler approaches have to be chosen
State of the art
– consider flexibility of the ligands, ignore that of the protein

Scoring functions for docking

Quantity of interest – binding free energy

But all of our free energy methods are too inefficient for docking!

Needed – extremely efficient way to quantify strength of binding

- to find the right binding mode of each ligand
- to compare the strength of binding of various ligands

Solution – scoring function

Scoring functions for docking

- based on force fields
  - Goldscore, DOCK, Autodock
- empirical
  - parametrized against experimental binding affinities
  - several chemically motivated contributions
  - ChemScore, PLP, Glide SP/XP
- knowledge-based
  - based on Boltzmann: frequently occuring motifs must have more negative binding free energy
  - PMF, DrugScore, ASP
- based on quantum-chemical calculations
  - using semi-empirical methods

Scoring functions for docking

$$\Delta G_{\text{bind}} = \Delta G_{\text{solvent}} + \Delta G_{\text{conf}} + \Delta G_{\text{int}} + \Delta G_{\text{rot}} + \Delta G_{\text{t/r}} + \Delta G_{\text{vib}}$$

$\Delta G_{\text{solvent}}$ – hydration changes during binding
ΔG_{conf} – conformation of the ligand – 'deformation energy' (binding pocket may constrain the ligand, and this costs energy)
ΔG_{int} – 'interaction energy' – specific interaction, favorable
ΔG_{rot} – loss of entropy ($\Delta G = -T \cdot \Delta S$) by frozen rotations around single bonds, approx. $+RT \log 3 = 0.7$ kcal/mol per rotatable bond with 3 states (trans, $2\times$ gauche)
$\Delta G_{\text{t/r}}$ – loss of trans+rot entropy upon association, approx. const.
ΔG_{vib} – change of vib. modes (entropy), difficult, often ignored

Scoring functions for docking

SF is actually a kind of force field for ΔG of binding

Problem – even though it is largely approximative,
it may be still computationally too costly to evaluate
for a huge number of ligands that is usually to be processed.

$\rightarrow$ the many SF proposed so far are usually extremely simple,
looking over-simplified in comparison with MM force fields.

Scoring functions for docking

An illustrative example (Böhm 1994):

$$\Delta G = \Delta G_0 + \Delta G_{\text{Hbond}} \cdot \sum_{\text{Hbonds}} f(R,\alpha) + \Delta G_{\text{ionpair}} \cdot \sum_{\text{ionpairs}} f'(R,\alpha) + \Delta G_{\text{lipo}} \cdot A_{\text{lipo}} + \Delta G_{\text{rot}} \cdot N_{\text{rot}}$$

ΔG_0 – constant term
ΔG_{Hbond} – ideal hydrogen bond
$f(R,\alpha)$ – penalty for realistic H bond (length R, angle α)
$\Delta G_{\text{ionpair}}$ and $f'(R,\alpha)$ – analogic quantities for ionic contacts
ΔG_{lipo} – from hydrophobic interaction,
proportional to the area of non-polar surface of molecule A_{lipo}
N_{rot} – # of rotatable bonds in ligand being frozen upon binding

Scoring functions for docking

Other development:

- partition surface areas of protein and ligand – polar/nonpolar, and assign different parameters to interactions
  polar-polar, polar-nonpolar, nonpolar-nonpolar
- statistical techniques to derive SF and its parameters

Problem of SF – only describes well tightly bound ligands
Weaker binding ligands (of increasing interest in docking studies)
– rather poorly described
– binding strength possibly overestimated – false positives
Possible solution – 'consensus scoring' – combine several SF

Note: error of ΔG of 1.4 kcal/mol – 10$\times$ error of binding constant
4.2 kcal/mol of ΔG lies between micro- and nanomolar inhibitor
– deadly difference – illustrates the requirements on accuracy

Some technical details

Grid

- observation: rigid receptor is identical for many calculations
- interactions with receptor can be pre-calculated on a grid
  – each calculation of SF is then accelerated

Preparation of the receptor

- can the active site relax (change structure) upon binding?
- if yes, maybe take structure from a complex (delete ligand)
- ensemble docking – use several receptor structures,
  finally choose the one to which the ligand binds most strongly
- flexible docking – (partially) flexible receptor, e.g. side chains
- identify protonation states of AAs (most difficult – His)

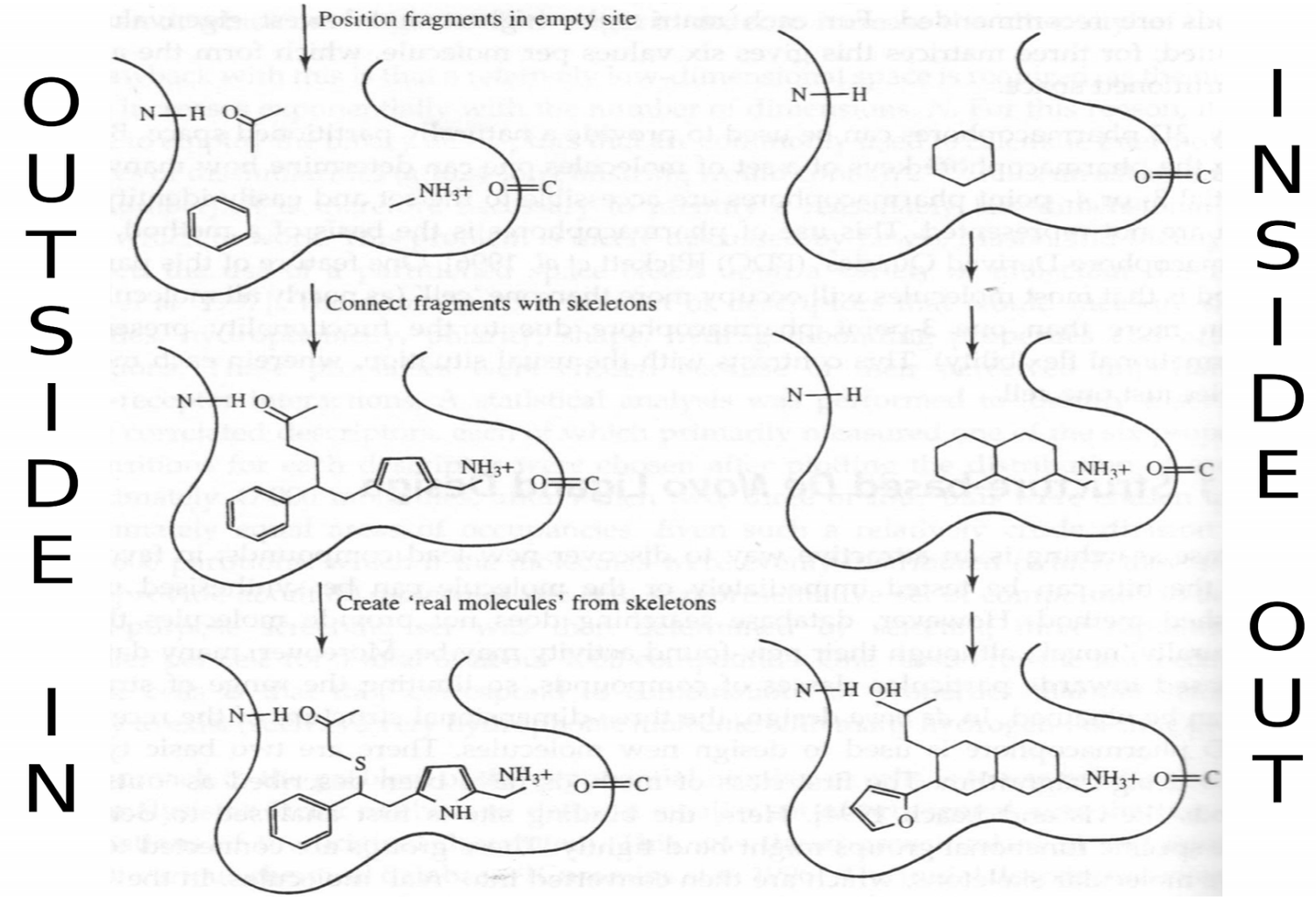
De novo design of ligands

It is very often useful to search a database of molecules,
but there is still a chance to miss the 'ideal' ligand
because no such compound is in the database

What about to construct the ligand 'from scratch'
– not relying on a database?

- 'outside–in' approach – binding analyzed first,
  and tightly-binding ligand fragments proposed,
  then connected together (database of linkers)
  $\rightarrow$ molecular skeleton
- 'inside–out' approach
  – 'growing' the ligand in the binding pocket,
  driven by a search algorithm with a scoring function

De novo design of ligands



De novo design of ligands

eHiTS – example of a novel approach