COMPUTATIONAL BIOCHEMISTRY

Ferenc Bogár

György Ferency

Eufrozina A. Hoffmann

Tamás Körtvélyesi

Eszter Németh

Gábor Paragí

Róbert Rajkó

COMPUTATIONAL BIOCHEMISTRY

Ferenc Bogár
György Ferency
Eufrozina A. Hoffmann
Tamás Körtvélyesi
Eszter Németh
Gábor Paragi
Róbert Rajkó

Publication date 2013.

TÁMOP-4.1.2.A/1-11/1 MSc Tananyagfejlesztés

Interdiszciplináris és komplex megközelítésű digitális tananyagfejlesztés a természettudományi képzési terület mesterszakjaihoz

Table of Contents

List of Tables

Preface

A jelen digitális tananyag a TÁMOP-4.1.2.A/1-11/1-2011-0025 számú, "Interdiszciplináris és komplex megközelítésű digitális tananyagfejlesztés a természettudományi képzési terület mesterszakjaihoz" című projekt részeként készült el.

A projekt általános célja a XXI. század igényeinek megfelelő természettudományos felsőoktatás alapjainak a megteremtése. A projekt konkrét célja a természettudományi mesterképzés kompetenciaalapú és módszertani megújítása, mely folyamatosan képes kezelni a társadalmi-gazdasági változásokat, a legújabb tudományos eredményeket, és az info-kommunikációs technológia (IKT) eszköztárát használja.

The *Computational Biochemistry* digital textbook was supported by the grant of TÁMOP-4.1.2.A/I-II/1-2011-0025. The developement of the Curricula was performed by professors and researchers accepted internationally by their research and publications from the University of Szeged, Hungarian Academy of Science, Chemaxon Ltd. and Semmelweis Medical School, Budapest.

No any digital textbooks are available for studying this exciting topic in Hungarian and in English. In English a lot of articles, textbooks and books are available which will be presented in the end of all Chapters as *Further Readings*. The topic of this digital textbook is suggested not only for chemists M.Sc., but all of the other natural science and technical M.Sc. students (biologist, biophysicists, physisicts, material scientists, environmental scientists, bioengineers, molecular biologists, bionics students).

The textbook includes thirteen chapters, which have *References*, *Further Readings* and *Questions* in the end of the chapters. A *Treasury of Theorems* supports the better understanding of the topics in the end of the book.

Chapter 1 includes the basic knowledge of the structure, intra- and intermolecular interactions in biologically active molecules (peptides, proteins, DNAs, PNAs, etc.). In Chapter 2 we summerized the simplest methods for the calculation of the structures in the molecules mentioned before. The biologically active molecules are working in solution, in water with interaction with the solvent molecules, ions and with each others. The solvation models are described in Chapter 3 and Chapter 4. Chapter 3 includes the implicit solvation models, the solution of the Poisson-Boltzmann equation which is one of the possibilities to predict the pK values of protonations in the charged side chains in peptide and proteins. In Chapter 4 the explicit solvent model and the implicit solvent models are described. Chapter 5 deals with the pK calculations of the side chains in peptides and proteins which is very important in the modelling of the structures in molecular mechanics and molecular dynamics calculations and in docking with ligand (drug-like) molecules. The basis of the molecular dynamics is summerized in Chapter 6. In some cases the experimental determination of the 3D protein structures has missing parts but with known sequence(s). There are some methods to predict the 3D structures which can be found in Chapter 7. The methods can be considered with critics. Computational methods for binding modes of protein-protein and protein-ligand (drug-like) molecules are detailed in Chapter 8. It includes the calculation of binding free energy and the methods of rescoring by empirical functions. Chapter 9 deals with the calculations of the binding free energies of drug-like molecules by molecular simulation/computational methods. The basic statistical methods are summerized in Chapter 10 as the introduction to the Cheminformatics. Quantum mechanics and mixed quantum mechanics/molecular mechanics (QM/MM) quantum mechanics/molecular dynamics (QM/MD) methods in the prediction of structure, intra- and intermolecular interactions and reactions of biologically active molecules can be found in Chapter 11. Chapter 12 summerizes the reaction kinetics of biological systems with the definitions. In Chapter 13 we try to give some case studies on the topics mentioned above.

Molecular graphics: Molegro Molecular Viewer 2.5, Molegro SA, www.molegro.com.

Molecular animation: ICM Browser Pro, icm-browser-3.7-2e-linux.sh, Molsoft LLC, San Diego USA.

Szeged, 17-05-2013

Tamas Kortvelyesi

associate professor

Department of Physical Chemistry and Material Science

University of Szeged

Chapters and Authors

Intra- and Intermolecular Interactions in Biologically Active Molecules. Structure of Peptides, Proteins, DNA and PNA. (Tamás Körtvélyesi)

Molecular Mechanics. (Tamás Körtvélyesi)

Electrostatics in Molecules. (Tamás Körtvélyesi)

Solvation Models. (Tamás Körtvélyesi)

Molecular Dynamics. (Ferenc Bogár)

Electrostatic in Molecules. (Tamás Körtvélyesi)

Prediction of Protein Structures and a Part of the Protein Structure. (Tamás Körtvélyesi)

Protein-protein and Protein-ligand Binding. Docking methods. (Tamás Körtvélyesi)

Calculation of Protein-Ligand Binding Free Energy. (György Ferenczy)

Introduction to Cheminformatics. Databases. (Róbert Rajkó, Tamás Körtvélyesi)

Quantum Mechanics and Mixed Quantum Mechanics/Molecular Mechanics Methods to Characterize the Structure and Reactions of Biologically Active Molecules.

(Gábor Paragi, György Ferenczy)

Evaluation of Reaction Kinetics Data. (Eufrozina A. Hoffmann)

Case Studies. Applications to biochemical problems. Some Examples on the Application of the Previous Computational Methods. (Gábor Paragi, Ferenc Bogár, Róbert Rajkó, György Ferenczy, Eufrozina A. Hoffmann, Tamás Körtvélyesi)

Further Readings in Hungarian

1. Keserű György Miklós, Kolossváry István, Molekulamechanika. A kémia legújabb eredményei 2003. Akadémia Kiadó, Budapest, 2003.

2. Keserű György Miklós, Kolossváry István, Bevezetés a számítógépes gyógyszertervezésbe. A kémia legújabb eredményei 2006. Akadémia Kiadó, Budapest, 2006.

3. A gyógyszerkutatás kémiája, Szerk. Keserű György Miklós, Akadémia Kiadó, Budapest, 2011.

Acknowledgment

Thank you for possibilities the sponsorship of

Chapter 1. Intra- and intermolecular interactions in biologically active molecules. structure of peptides, proteins, dna and pna

(Tamás Körtvélyesi)

Keywords: biomolecules, salt bridges, H-bonds, π-π stackings, π-HX stackings, peptide bond, protein, primary-, secondary-, tertiary-, quatarnary-structure

What is described here? The intra- and intermolecular interactions which stabilize the molecular structures are summerized. No deteiled description of the classes of biomolecules are described here, because a lot of books (and Biochemistry course) deal with this topic (see e.g. the Further Readings in English and in Hungarian). We do not strive for the complete discussion of the biomolecules and biochemical mechanisms.

What is it used for? The knowledge is important for the mathematical description of the interactions – the necessarry expressions and the neglection of some expressions without great errors.

What is needed? The basic knowledge in organic chemistry, biochemistry and physical chemistry is necessarry.

1. Introduction

The main compounds in the living cells are peptides, proteins, lipids, sugars, phospholipids, DNA, water and salts, etc. with different functions (structural molecules, enzymes, etc.). These molecules and ions are important in working of living cells. The main interactions in biomolecules are described in this chapter. Some of the biomolecules are depicted by animations to study the building blocks of the molecules.We do not show the molecules, the classes of molecules, because the reader can find these information in a lot of excellent textbooks.

2. Intramolecular Interactions Stabilizing the Structure

2.1. Peptide bonds

The 20 native aminoacids (with L-chirality) can bind to each other by peptide bonds. The four-atom link is called peptide link (http://en.wikipedia.org/wiki/Peptide_bond). Peptide bonds are mainly trans peptide binding of residues. The –C(=O)NH is a resonance stabilized structure which is planar. It is sensitive to water and pH, and they can break easily (that is why the peptides can not be used as drugs through mouth). In some cases – mainly at Pro the ratio of the cis/trans isomers is 1:3. (It is a good possibility for the validation of MD.)

„Within three, four or five residues, the turns are assigned on the basis of the position of the H-bond between the residues, *i* to *i+2* to *i+3* and *i* to *i+4*. A turn is marked at position *i* to *i+1* for the three residue turn, *i+1* to *i+2* for a four residue turn, and *i+1*, *i+2*, *i+3* for a five residue turn. A β-bridge is assigned when two non overlapping stretches of three residues each, *i-1*,*i*,*i+1* and *j-1*,*j*,*j+1* form a hydrogen bonding pattern consistent with either parallel or antiparallel β-structures and is marked at the *i* and *j* residues. β-sheet is defined by more consecutive β-bridges. The bend is defined as a five residue turn without H bondsTwo consecutive turns at position *i-1*and *i* form a 3_{10}-helix which is marked at *i*, *i+1* and *i+2*. α-helix at *i*, *i+1*, *i+3* and π-helix at *i*, *i+1*, *i+2*, *i+3*, *i+4*" [1,2].

2.2. Salt bridges

Strong electrostatic interactions are due to the attraction of positive-negative charges of groups or the repulsion of the same charges. Thy have great effect on t he stability of the structures and ont he formation of secondary,

tertiary and quatarnary structures in proteins and in the other biomolecules (see later) [1]. The electrostatic interactions can be calculated by the Coulomb equation (see Chapter 1). The Coulomb interactions are not only between the groups with integer charges (Lys, Asp, Glu, Arg, etc.), but between groups with partial charges (see Lit. [2,3]).

2.3. H-Bonds

Pauling [4] suggested a secondary bond which can be characterized by X-H…Y where X and Y as pilar atoms have greater electronegativity [3]. The energy stability is increasing and the geometries are deformed. F−H…:F (161.5 kJ/mol), O−H…:N (29 kJ/mol), O−H…:O (21 kJ/mol), N−H…:N (13 kJ/mol), N−H…:O (8 kJ/mol), HO−H…:OH (18 kJ/mol). X−H…Y system: X−H distance is typically ca. 110 pm, http://en.wikipedia.org/wiki/Picometre \o "Picometre" whereas H…Y distance is ca.160 to 200 pm. As it can be seen, the stabilization energies are much more lower than the chemical bonds. Bondi suggested a geometrical description of H-bonds [5]. H-bond can be defined by the van der Waals radiuses and angles. If the distance of the pilar atoms (X, Y) is less than the sum of the van der Waals radii and the angle of X-H…Y is greater than 90 degree (and less than 180 degree). The best computational method to recognize the H-bonds is DSSP [6] which was developed by Kabsch and Sanders [7].

2.4. π- π, π-HN, π-HO and π-H_3N^+ Stacking

π-π (see *Figure 1.1*), π-HN, π-OH stackings are weakly polar interactions (π-π interaction ca. 5-10 kJ/mol, the distance is ca. maximum 8 Á, π-HN interaction ca. 3-5 kJ/mol, the distance is ca. maximum 5 Á, π-OH stackings interaction ca. 3-5 kJ/mol, the distance is ca. maximum 5 Á) where a delocalized π-electron system (dipole, quadrupole) interact with other π-delocalized electron system, HN- and, HO-. Petzko et al. [8-11] proved that these interactions have significant effect on the structure of peptides and proteins. The effect of π-HN on the peptide structures were supported by molecular dynamics (MD) [2]. A rethinking of π-stacking interactions are published recently [13]. An important interaction is the π-H_3N^+. The negatively charged delocalized π-system electrostatically interact with the positively charge N-terminal or positively charged Lys side chain (or positively charged Arg).

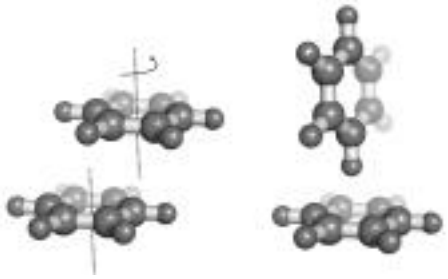


Figure 1.1. π-π stacking between aromatic groups

2.5. Hydrophobic Interactions

Interactions with dispersion are important between molecules (e.g. alkanes) without charges [13]. The alkane/water and octanol/water partition coefficients are important in drug design to predict the solution of drug molecules in the cells. The possibility of the predictions of the partition coefficients are summerized in Lit. [13].

2.6. Protein-metal complexes

A lot of metals can bind to the proteins (Zn, Fe, Co, Mg, etc). They have basic roles in the biochemical reactions (see e.g. Zn-fingers, heamoglobine, etc.).

.

3. Peptides and Proteins Structures

The 20 native aminoacids are summerized in *Figure 1.2*. on the basis of the polarity of the side-chains (hydrophobic, polar, acidic, basic) with the three-letters and one-letter codes.

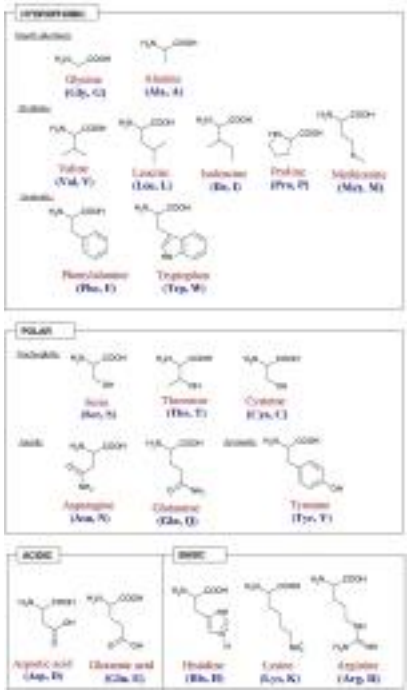


Figure 1.2. Natural aminoacids classified by the polarity of the side-chain

Figure 1.3. summerizes the sequence (primary structure) of βA(1-40) amyloid peptide structure (PDB Id.: 1aml). In *Figure 1.4*. the peptide is described with polar H-atoms with secondary structure.

```
SEQRES   1 A   40  ASP ALA GLU PHE ARG HIS ASP SER GLY TYR GLU VAL HIS
SEQRES   2 A   40  HIS GLN LYS LEU VAL PHE PHE ALA GLU ASP VAL GLY SER
SEQRES   3 A   40  ASN LYS GLY ALA ILE ILE GLY LEU MET VAL GLY GLY VAL
SEQRES   4 A   40  VAL
```

Figure 1.3. βA(1-40) amyloid peptide structure with sequence (PDB Id.: 1aml)

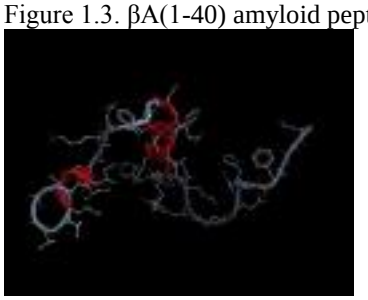


Figure 1.4. Animation of βA(1-40) amyloid peptide structure with polar H atoms

The sequence of the chain A (of the four chains) in β-secretase (PDB Id.: 1fkn, menapsin 2, Homo Sapiens) can be seen in *Figure 1.5*. The chain A in β-secretase (PDB Id.: 1fkn) is described in *Figure 1.6*. with side chains without H-atoms and with the secondary structures.

```
SEQRES   1 A  391  ARG ARG GLY SER PHE VAL GLU MET VAL ASP ASN LEU ARG
SEQRES   2 A  391  GLY LYS SER GLY GLN GLY TYR TYR VAL GLU MET THR VAL
SEQRES   3 A  391  GLY SER PRO PRO GLN THR LEU ASN ILE LEU VAL ASP THR
SEQRES   4 A  391  GLY SER SER ASN PHE ALA VAL GLY ALA ALA PRO HIS PRO
SEQRES   5 A  391  PHE LEU HIS ARG TYR TYR GLN ARG GLN LEU SER SER THR
SEQRES   6 A  391  TYR ARG ASP LEU ARG LYS GLY VAL TYR VAL PRO TYR THR
SEQRES   7 A  391  GLN GLY LYS TRP GLU GLY GLU LEU GLY THR ASP LEU VAL
SEQRES   8 A  391  SER ILE PRO HIS GLY PRO ASN VAL THR VAL ARG ALA ASN
SEQRES   9 A  391  ILE ALA ALA ILE THR GLU SER ASP LYS PHE PHE ILE ASN
SEQRES  10 A  391  GLY SER ASN TRP GLU GLY ILE LEU GLY LEU ALA TYR ALA
SEQRES  11 A  391  GLU ILE ALA ARG PRO ASP ASP SER LEU GLU PRO PHE PHE
SEQRES  12 A  391  ASP SER LEU VAL LYS GLN THR HIS VAL PRO ASN LEU PHE
SEQRES  13 A  391  SER LEU GLN LEU CYS GLY ALA GLY PHE PRO LEU ASN GLN
SEQRES  14 A  391  SER GLU VAL LEU ALA SER VAL GLY GLY SER MET ILE ILE
SEQRES  15 A  391  GLY GLY ILE ASP HIS SER LEU TYR THR GLY SER LEU TRP
SEQRES  16 A  391  TYR THR PRO ILE ARG ARG GLU TRP TYR TYR GLU VAL ILE
SEQRES  17 A  391  ILE VAL ARG VAL GLU ILE ASN GLY GLN ASP LEU LYS MET
SEQRES  18 A  391  ASP CYS LYS GLU TYR ASN TYR ASP LYS SER ILE VAL ASP
SEQRES  19 A  391  SER GLY THR THR ASN LEU ARG LEU PRO LYS LYS VAL PHE
SEQRES  20 A  391  GLU ALA ALA VAL LYS SER ILE LYS ALA ALA SER SER THR
SEQRES  21 A  391  GLU LYS PHE PRO ASP GLY PHE TRP LEU GLY GLU GLN LEU
SEQRES  22 A  391  VAL CYS TRP GLN ALA GLY THR THR PRO TRP ASN ILE PHE
SEQRES  23 A  391  PRO VAL ILE SER LEU TYR LEU MET GLY GLU VAL THR ASN
SEQRES  24 A  391  GLN SER PHE ARG ILE THR ILE LEU PRO GLN GLN TYR LEU
SEQRES  25 A  391  ARG PRO VAL GLU ASP VAL ALA THR SER GLN ASP ASP CYS
SEQRES  26 A  391  TYR LYS PHE ALA ILE SER GLN SER SER THR GLY THR VAL
SEQRES  27 A  391  MET GLY ALA VAL ILE MET GLU GLY PHE TYR VAL VAL PHE
SEQRES  28 A  391  ASP ARG ALA ARG LYS ARG ILE GLY PHE ALA VAL SER ALA
SEQRES  29 A  391  CYS HIS VAL HIS ASP GLU PHE ARG THR ALA ALA VAL GLU
SEQRES  30 A  391  GLY PRO PHE VAL THR LEU ASP MET GLU ASP CYS GLY TYR
SEQRES  31 A  391  ASN
```

Figure 1.5. β-secretase (PDB Id.: 1fkn, menapsin 2, Homo Sapiens) sequence

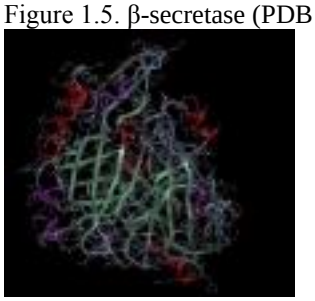


Figure 1.6. Animation of β-secretase chain A (PDB Id.: 1fkn, menapsin 2, Homo Sapiens)

```
SEQRES   1 A  110  ALA GLN VAL ILE ASN THR PHE ASP GLY VAL ALA ASP TYR
SEQRES   2 A  110  LEU GLN THR TYR HIS LYS LEU PRO ASP ASN TYR ILE THR
SEQRES   3 A  110  LYS SER GLU ALA GLN ALA LEU GLY TRP VAL ALA SER LYS
SEQRES   4 A  110  GLY ASN LEU ALA ASP VAL ALA PRO GLY LYS SER ILE GLY
SEQRES   5 A  110  GLY ASP ILE PHE SER ASN ARG GLU GLY LYS LEU PRO GLY
SEQRES   6 A  110  LYS SER GLY ARG THR TRP ARG GLU ALA ASP ILE ASN TYR
SEQRES   7 A  110  THR SER GLY PHE ARG ASN SER ASP ARG ILE LEU TYR SER
SEQRES   8 A  110  SER ASP TRP LEU ILE TYR LYS THR THR ASP HIS TYR GLN
SEQRES   9 A  110  THR PHE THR LYS ILE ARG
SEQRES   1 B  110  ALA GLN VAL ILE ASN THR PHE ASP GLY VAL ALA ASP TYR
SEQRES   2 B  110  LEU GLN THR TYR HIS LYS LEU PRO ASP ASN TYR ILE THR
SEQRES   3 B  110  LYS SER GLU ALA GLN ALA LEU GLY TRP VAL ALA SER LYS
SEQRES   4 B  110  GLY ASN LEU ALA ASP VAL ALA PRO GLY LYS SER ILE GLY
SEQRES   5 B  110  GLY ASP ILE PHE SER ASN ARG GLU GLY LYS LEU PRO GLY
SEQRES   6 B  110  LYS SER GLY ARG THR TRP ARG GLU ALA ASP ILE ASN TYR
SEQRES   7 B  110  THR SER GLY PHE ARG ASN SER ASP ARG ILE LEU TYR SER
SEQRES   8 B  110  SER ASP TRP LEU ILE TYR LYS THR THR ASP HIS TYR GLN
SEQRES   9 B  110  THR PHE THR LYS ILE ARG
SEQRES   1 C  110  ALA GLN VAL ILE ASN THR PHE ASP GLY VAL ALA ASP TYR
SEQRES   2 C  110  LEU GLN THR TYR HIS LYS LEU PRO ASP ASN TYR ILE THR
SEQRES   3 C  110  LYS SER GLU ALA GLN ALA LEU GLY TRP VAL ALA SER LYS
SEQRES   4 C  110  GLY ASN LEU ALA ASP VAL ALA PRO GLY LYS SER ILE GLY
SEQRES   5 C  110  GLY ASP ILE PHE SER ASN ARG GLU GLY LYS LEU PRO GLY
SEQRES   6 C  110  LYS SER GLY ARG THR TRP ARG GLU ALA ASP ILE ASN TYR
SEQRES   7 C  110  THR SER GLY PHE ARG ASN SER ASP ARG ILE LEU TYR SER
SEQRES   8 C  110  SER ASP TRP LEU ILE TYR LYS THR THR ASP HIS TYR GLN
SEQRES   9 C  110  THR PHE THR LYS ILE ARG
SEQRES   1 D   89  LYS LYS ALA VAL ILE ASN GLY GLU GLN ILE ARG SER ILE
SEQRES   2 D   89  SER ASP LEU HIS GLN THR LEU LYS LYS GLU LEU ALA LEU
SEQRES   3 D   89  PRO GLU TYR TYR GLY GLU ASN LEU ASP ALA LEU TRP ASP
SEQRES   4 D   89  ALA LEU THR GLY TRP VAL GLU TYR PRO LEU VAL LEU GLU
SEQRES   5 D   89  TRP ARG GLN PHE GLU GLN SER LYS GLN LEU THR GLU ASN
SEQRES   6 D   89  GLY ALA GLU SER VAL LEU GLN VAL PHE ARG GLU ALA LYS
SEQRES   7 D   89  ALA GLU GLY ALA ASP ILE THR ILE ILE LEU SER
SEQRES   1 E   89  LYS LYS ALA VAL ILE ASN GLY GLU GLN ILE ARG SER ILE
SEQRES   2 E   89  SER ASP LEU HIS GLN THR LEU LYS LYS GLU LEU ALA LEU
SEQRES   3 E   89  PRO GLU TYR TYR GLY GLU ASN LEU ASP ALA LEU TRP ASP
SEQRES   4 E   89  ALA LEU THR GLY TRP VAL GLU TYR PRO LEU VAL LEU GLU
SEQRES   5 E   89  TRP ARG GLN PHE GLU GLN SER LYS GLN LEU THR GLU ASN
SEQRES   6 E   89  GLY ALA GLU SER VAL LEU GLN VAL PHE ARG GLU ALA LYS
SEQRES   7 E   89  ALA GLU GLY ALA ASP ILE THR ILE ILE LEU SER
SEQRES   1 F   89  LYS LYS ALA VAL ILE ASN GLY GLU GLN ILE ARG SER ILE
SEQRES   2 F   89  SER ASP LEU HIS GLN THR LEU LYS LYS GLU LEU ALA LEU
SEQRES   3 F   89  PRO GLU TYR TYR GLY GLU ASN LEU ASP ALA LEU TRP ASP
SEQRES   4 F   89  ALA LEU THR GLY TRP VAL GLU TYR PRO LEU VAL LEU GLU
SEQRES   5 F   89  TRP ARG GLN PHE GLU GLN SER LYS GLN LEU THR GLU ASN
SEQRES   6 F   89  GLY ALA GLU SER VAL LEU GLN VAL PHE ARG GLU ALA LYS
SEQRES   7 F   89  ALA GLU GLY ALA ASP ILE THR ILE ILE LEU SER
```

Figure 1.7. The sequence of Barnase-Barstar protein with chains A, B, C, D, E, F (PDB Id.: 1brs)

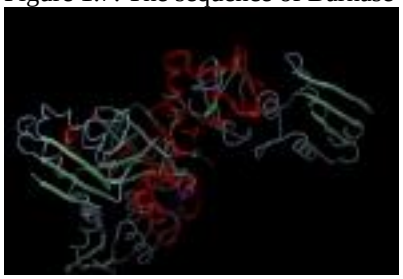


Figure 1.8. Animation of the secondary and tertiary structure of Barnase-Barstar protein with chains A, B, C, D, E, F (PDB Id.: 1brs)

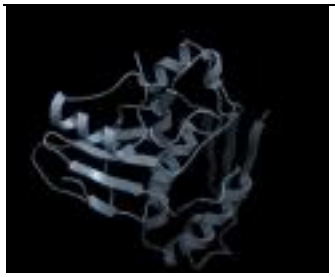


Figure 1.9. Animation of the secondary structure of Barnase-Barstar protein with chains A and D (PDB Id.: 1b2u)

4. DNA and PNA Structures

The primary structure of a linear sequence of nucleotides linked with phoshodiester bonds. The nucleic acid sequence is the primary structure (*Figure 1.10*.). The secondary structure depends on the H-bonds and stacking interactions between the basis (*Figure 1.11*.). The two rings in adenine and the 5-membered ring in guanine is aromatic. Neither cytosine and thymine are aromatic (http://en.wikipedia.org/wiki/Nucleic_acid_structure). The tertiary structure can be in DNA's double helix B-DNA, A-DNA and Z-DNA. The quatarnary structure of DNA is similar to that of protein: the interactions with proteins and other DNAs (*Figure 1.12*. and *1.13*.).

PNAs were developed int he early 90s. In DNA, RNA the backbone is sensitive to pH because of the charged phosphate backbone. In PNA the backbone was changed to a neutral, geometrically almost the same backbone. This bacbone is not sensitive to pH. It is a good possibility to develope DNA-chips to recognize the basis in DNA. It is also a good possibility for the personal therepeautic procedure (see e.g. *Figure 1.15*.).

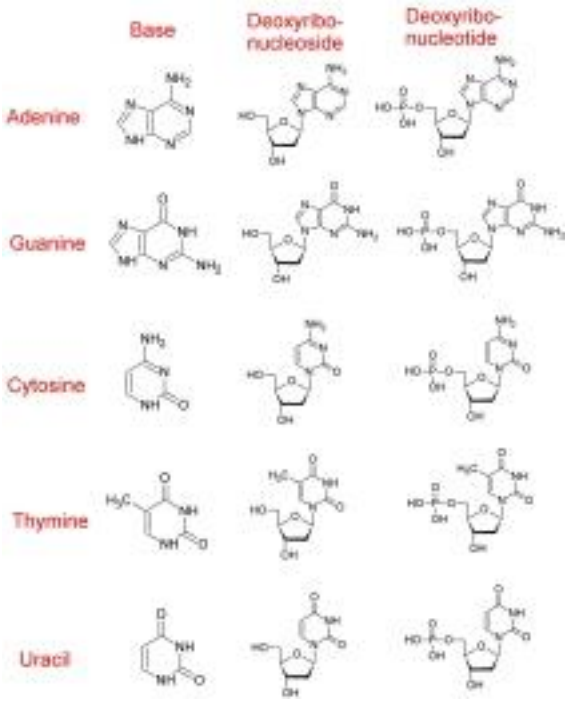


Figure 1.10. The nucleinbasis, deoxyriboze-nucleosides, deoxyriboze-nucleotides (A, G, C, T, U basis.

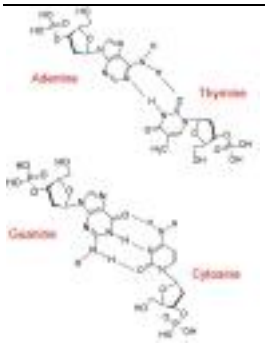


Figure 1.11. The Watson-Crick duplexes with H-bonds of AMP::TMP and GMP::CMP

```
SEQRES   1 A   14   DG  DC  DG  DC  DA  DT  DG  DC  DT  DA  DC  DG  DC
SEQRES   2 A   14   DG
SEQRES   1 B   14   DC  DG  DC  DG  DT  DA  DG  DC  DA  DT  DG  DC  DG
SEQRES   2 B   14   DC
```

Figure 1.12. Sequence of duplex DNA (PDB Id.: 2m2c)

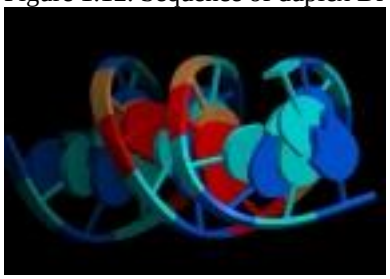


Figure 1.13. Animation of duplex DNA Chain A: (5'-D(*GP*CP*GP*CP*AP*TP*GP*CP* TP*AP*CP*GP*CP*G)-3'); Chain B: (5'-D(*CP*GP*CP*GP*TP*AP*GP*CP*AP*TP* GP*CP*GP*C)-3'). (blue: G,C, red:A, T) (PDB Id.: 2m2c, NMR structure)

```
A, C: (5'-D(*CP*(BRU)P*CP*CP*(BRU), 3 P*CP*CP*GP*CP*GP*CP*G)-3');
B, D: (5'-D(*CP*GP*CP*GP*CP*GP*GP*AP*G)-3');
```

Figure 1.14. The sequence of triplex DNA and its junction with a duplex DNA (XRD result) (PDB Id.: 1d3r)

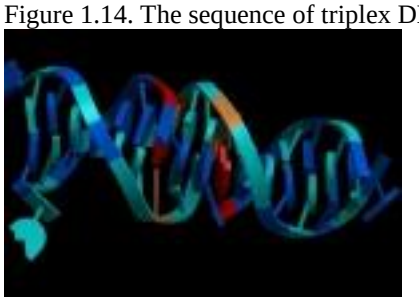


Figure 1.15. The animation of triplex DNA and its junction with a duplex DNA (blue: G, C, red:A, T) (XRD result) (PDB Id.: 1d3r)

```
SEQRES   1 A    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 B    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 C    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 D    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 E    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 F    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 G    7  CP1 GPN TP1 APN CP1 GPN LYS
SEQRES   1 H    7  CP1 GPN TP1 APN CP1 GPN LYS
```

Figure 1.16. The sequence of PNA duplexes (Hexamer, Chains A, B, C, D, E, F, G, H) (XRD result) (PDB Id.: 1qpy)

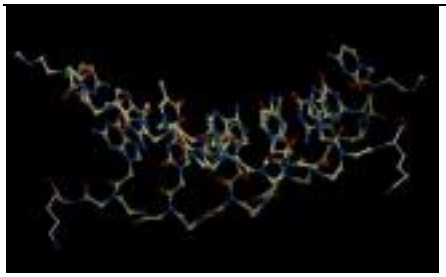


Figure 1.17. The animation of PNA duplex (Chain A and B) without H-atoms (Double stranded helix, P-form, right and left handed helix) (XRD result) (PDB Id.: 1qpy)

```
SEQRES    1 A   24  CPN TPN CPN TPN IPN CPN TPN TPN CPN HIS GLY SER SER
SEQRES    2 A   24  GLY HIS CPN TPN TPN CPN TPN TPN CPN TPN CPN
SEQRES    1 B    9   DG  DA  DA  DG  DA  DA  DG  DA  DG
SEQRES    1 C   24  CPN TPN CPN TPN IPN CPN TPN TPN CPN HIS GLY SER SER
SEQRES    2 C   24  GLY HIS CPN TPN TPN CPN TPN TPN CPN TPN CPN
SEQRES    1 D    9   DG  DA  DA  DG  DA  DA  DG  DA  DG
```

Figure 1.18. The sequence of hairpin PNA/DNA triplex (XRD result) (PDB Id.: 1pnn)

The H-bonds based on Watson-Crick and Hoogsten type interactions.

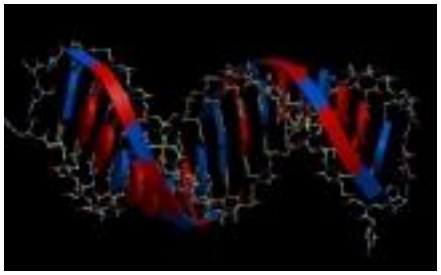


Figure 1.19. The animation of PNA/DNA triplex without H-atoms (blue: G, red:A) (XRD result) (PDB Id.: 1pnn)

5. Membranes

Membranes are important int he structure of cells. Membrane proteins are necessarry in the ion channels in the cells. They are built up by phospho lipids. Some of the main phospholipid molecules are DPC (dodecylphosphocholine), DPPC (dipalmitoylphosphatidyl-choline), DPMC (1,2-dimyristoyl-sn-glycero-3-phosphocholine), POPC ([(2R)-3-hexadecanoyloxy-2-[(Z)-octadec-9-enoyl]oxypropyl] 2-(trimethylazaniumyl)ethyl phosphate), POPE (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine). The animations of the structures (available in http://people.ucalgary.ca/~tieleman/download.html) are in *Figure 1.20.* to *1.24.*)

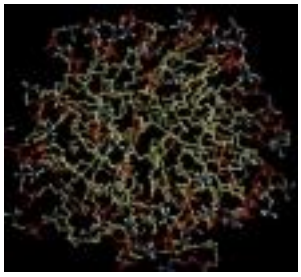


Figure 1.20. The animation of a micella (65 DPC lipids without water molecules)

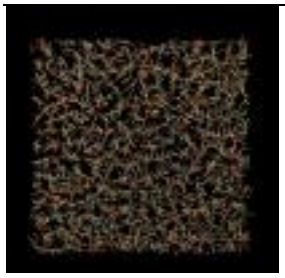


Figure 1.21. The animation of DPPC (128 DPPC without water)

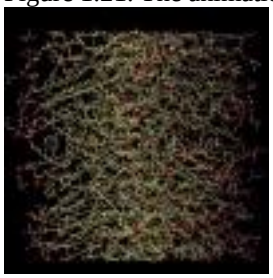


Figure 1.22. The animation of DMPC (128 DMPC without water)

Figure 1.23. The animation of POPC (128 POPC without water)

Figure 1.24. The animation of POPE (128 POPE without water)

6. Databases

The experimental structures of biomolecules are deposited in http://www.pdb.org. These structures are the results of XRD and NMR experiments and freely available. Presently, ca. 90424 3D structures (with redundantly) (05.10.2013) are available in the databank. Data includes the 3D structures, experimental details, citation, etc.

ExPASy is the SIB Bioinformatics Resource Portal (http://www.expasy.org/) which provides access to scientific databases and software tools etc. in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. The database inludes the sequence of the proteins, and some larger systems.. No 3D structures are available.

There are special databasis which are commercial and the includes ca. 3.5-4 million compounds (e.g. Available Chemical Database – ACD, Accelrys). Some companies dealing with special fine chemicals suggest compounds on internet or CD with 3D structures and chemical properties (Mayflower, Asinex, etc.). These databases are

available for finding the best scaffold in docking (for validation see e.g. Lit. [15]). The ligand molecules checked biologically for cancer and HIV can be found in NCI (National Cancer Institute of NIH) [16].

7. Summary

The main classes of biomolecules were described without completeness. The main databasis which include the biomolecules and ligand structures were also described.

8. References

1. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637(1983).

2. T. Körtvélyesi, R. F. Murph y, S.Lovas/ , Secondary structures and intramolecular interactions in fragments of the B-loops of naturally occurring analogs of epidermal growth factor. J. of Biomol. Struct. & Dyn. 17(2):393-407(1999).

3. H. R. Bosshard, D. N. Marti, I. Jelesarov, Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. J. Mol. Recognit. 17(1):1-16(2004)..

4. S. Costantini, G. Colonna, A. M. Facchiano ESBRI: A web server for evaluating salt bridges in proteins. Bioinformation 3(3): 137–138(2008).

5. http://bioinformatica.isa.cnr.it/ESBRI/

6. P. Linus, The Nature of the Chemical Bond. Ithaca, NY: Cornell University Press, 1945.

7. A. Bondi, A., Van der Waals Volumes and Radii. J. Phys. Chem. 68 (3), 441-451(1964).

8. R. P. Joosten, T. A. H. Te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, G. Vriend, A series of PDB related databases for everyday needs., NAR 2010; doi: 10.1093/nar/gkq1105.

9. S. K. Burley, G. A. Petsko, Amino-aromatic interactions in proteins. FEBS Lett. 28;203(2):139-43(1986).

10. S. K. Burley, G. A. Petsko, Aromatic-aromatic interaction: a mechanism of protein structure stabilization. Science 229(4708):23-8(1985).

11. S. K. Burley, G. A. Petsko, Weakly polar interactions in proteins. Adv. Protein Chem. 39, 125-89(1988).

12. G. A. Petsko, Analyzing molecular interactions. Curr. Protoc. Bioinformatics. Chapter 8:Unit8.1(2003).

13. C. R. Martinez, B. L. Iverson, Rethinking the term “pi-stacking, Chem. Sci., 2012,3, 2191-2201(2012).

14. Hydrophobicity and Solvation in Drug Design. Part III. Ed. Y. C. Martin, KLUWER/ESCOM, Perspectives in Drug Discovery and Design, Vol. 19, 2000.

15. a) J. J. Irwin, B. K. Shoichet, ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. J Chem Inf Model 45 (1), 177-82 (2005). b) http://zinc.docking.org/

16. http://cactus.nci.nih.gov/ncidb2.1/

9. Further Readings

1. C. K. Mathews, K. E. van Holde, K. G. Ahern, Biochemistry, Addison Wesley Longman, Inc., San Francisco, Reading, MA, New York, Harlow, England, Don Mills, ON, Sydney, Madrid, Amsterdam, 2000.

2. G. Zubay, Biochemistry, Wm. C. Brown Publishers, Third Edition, 1993.

3. Conformational Proteomics of Macromolecular Architecture. Approaching the Structure of Large Molecular Assemblies and Their Mechanisms of Action.Eds. R. Holland and L. Hammar, World Scientific, New Jersey, London, Singapore, Beijing, Shanghai, Hong Kong, Taipei, Chennai, World Scientific Publishing Co. Pte Ltd. 2004.

4. A. M. Lesk, Introduction to Protein Architecture, Oxford University Press,Oxford, New York, 2001.

5. A. Fersht, Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding.W.H. Freeman and Company, New York, 3rd Printing, 2000.

6. Novák L., Nyitrai J., Hazai L., Biomolekulák Kémiája, MKE, Budapest 2001.

7. Hollósi M., Laczkó I., Asbóth B., Biomolekuláris kémia I. Nemzeti Tankönyvkiadó, Budapest 2005.

8. Hollósi M., Asbóth B., Biomolekuláris kémia II. Nemzeti Tankönyvkiadó, Budapest 2007.

10. Questions

1. Please, describe the intra- and intermolecular interactions between non-charged groups!

2. Please, describe the intra- and intermolecular interactions between point-point, point-dipole and dipole-dipole charged groups (physical chemistry)!

3. Please, describe the geometrical description of H-bond by Bondi!

4. Please, describe the resonance stabilized peptide bonds!

5. Please, describe the π- π interactions!

6. Please, describe the π-H-O and π-H-N interactions!

11. Glossary

π-π , π-HN, π-HO stackings have interaction between aromatic electron systems and C-H, O-H, N-H atoms. These interactions are dipole, quadrupole etc. interactions.

H-bonds X-H…Y, where X and Y as pilar atoms have greater electronegativity. The energy stability is increasing and the geometries are deformed. F−H…:F (161.5 kJ/mol), O−H…:N (29 kJ/mol), O−H…:O (21 kJ/mol), N−H…:N (13 kJ/mol), N−H…:O (8 kJ/mol), HO−H…:OH (18 kJ/mol). X−H…Y system: X−H distance is typically ca. 110 pm, whereas H…Y distance is ca.160 to 200 pm.

Databases, The 3D structures (XRD, NMR) of biomolecules are deposited in http://www.pdb.org, The known sequeces of the living systems are summerized in http://www.expasy.org/.

Chapter 2. Molecular Mechanics

(Tamás Körtvélyesi)

Keywords: molecular mechnics methods, potential functions, deformations in molecules, polarizable molecular mechanics, force fields

What is described here? This chapter deals with the simplest method to calculate the structures and thermochemistry of organic compounds with special regards to biomolecules. The methods are available for the large molecular systems (with 100-200 thousands atoms) and the basis of the molecular dynamics calculations (see Chapter 5) and molecular docking methods (Chapter 8).

What is it used for? To optimize the geometries of molecules built up, conformational analysis, applied in molecular dynamics as potential functions, solvation thermodynamics, to find the energetics of intra- and intermolecular interactions, score function of docking (drug-

like) ligands to target(s).

What is needed? The knowledge of the structure of biomolecules, intra- and intermolecular interactions to stabilize their structure (covalent, polar and weakly polar interactions) are fundamental to understand this chapter. The basic knowledge in physics and physical chemistry is also important.

1. Introduction

In the early forties of the last century Westheimer, an organic chemist suggested a molecular model: the atom sin the molecule is connected by springs. The structure can be calculated considering the force constants of the springs between the atoms and the non-bonded interactions. The idea is simple but no computers were available at that time. Only in the fifties-sixties were developed algorithms which use of the idea mentioned above. The method is simple and fast to find the conformations, intra- and intermolecular interactions, electrostatic properties of small molecules and large molecules, too. Molecular mechanics is the only method to handle large (bio)molecular systems with 100-200 thousands of atoms [1]. The algorithm makes possible to apply in computer assisted drug design.

2. Traditional Molecular Mechanics Methods

On a multidimensional Born-Oppenheimer surface the nuclear positions are given by the function described in an Eq. 2.1:

$$\min \; V = f(x, y, z) \tag{2.1}$$

The potential energy function in a molecule can be partitioned by the deformations Eq. 2.2

$$V = V_{stretching} + V_{bending} + V_{torsion} + V_{oop} + V_{cross} + V_{nb} \tag{2.2}$$

where $V_{stretching}$ is potential of the deformation in the bonds, $V_{bending}$ is the potential of the deformation in the angle bending and $V_{torsion}$ is the potential of the deformation in the torsion angle (covalent deformation), V_{oop} is the out-of-plane deformation, V_{cross} is the cross functions of the covalent interactions . See *Figure 2.1*. V_{nb} is the potential energy of the non-bonding (the Coulomb and the van der Waals) interactions [1]. This sum is called steric energy.

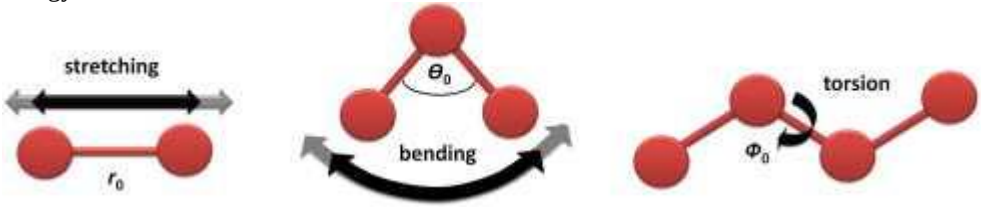


Figure 2.1. Deformations in molecular mechanics handled by the all-atom and united atom models (see later).

Covalent bonded interactions

The most simple function to describe the bond deformations is given by the Hooke's law Eq. 2.3.

$$V_{stretching}=\frac{1}{2}\sum_i k_{i,stretching}(r_i-r_{i,o})^2 \tag{2.3}$$

This function (which is suitable for small deformations, a harmonic vibrational potential) with the parameters of $k_{i,stretching}$ (stretching force constants) and $r_{i,o}$, natural bond length (the bond distances in an ideal strain free bond) for the individual bonds. E.g. $k_{i,stretching}$ is 272 kJ/mol. $r_{Csp3\text{-}Csp3,o}$ is 1.54 Å. The best $r_{i,o}$ values can be obtained by electron diffraction. At X-ray diffraction, the bond length of C-H has to correct by 0.015 Å because of the electronegativity difference between C and H. The Hooke's law is valid for small deformations, at strained systems the function is not precise. The equation can be modified as Eq. 2.4

$$V_{stretching}=\frac{1}{2}\sum_i k_{i,stretching}(r_i-r_{i,o})^2\left[1-k'_{i,stretching}(r_i-r_{i,o})-k''_{i,stretching}(r_i-r_{i,o})^2-\ldots\right]. \tag{2.4}$$

$k_{stretching}$, $k'_{stretching}$, $k''_{stretching}$ are the force constants.

In a strained molecular system the Morse function (the potential of the bond deformation in singlet state with dissociation) can be used Eq. 2.5 (see *Figure 2.2*).

$$V_{stretching}=\sum_i D_{e,i}(1-\exp(-\alpha_i X_i))^2 \tag{2.5}$$

$D_{e.i}$is the dissociation constant of the *i*th bond. α_i and X_i are the Morse constants of the bonds and the deformation of the bonds ($X_i = r - r_{i,o}$), respectively. The Morse constant of a bond pair can be calculated by

$$\alpha_i=\sqrt{\frac{k_{e,i}}{2D_{e,i}}} \tag{2.6}$$

where $k_{e,i}$ is the force constant of the *i*th bond.

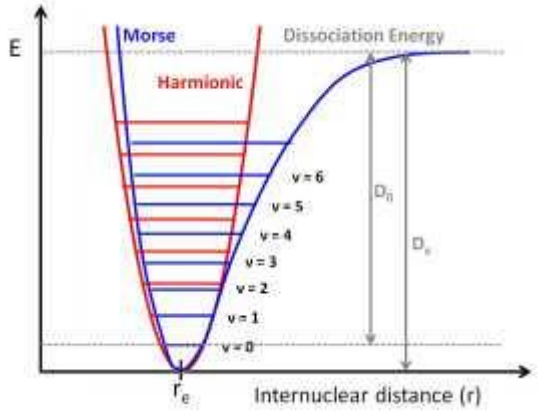


Figure 2.2. Dissociation energy profiles: harmonic and the energy profiles by Morse function. D_e is the dissociation energy, $D_0$is the ZPVE (zero point vibrational energy, $D_0 = D_e + ZPVE$, $v= 0$) corrected dissociation energy, r_e is the equilibrium distance, v is the energy level

The simplest function of the deformation in the angle can be described by Eq. 2.7

$$V_{bending}=\frac{1}{2}\sum_i k_{i,bending}(\theta_i-\theta_{i,o})^2 \tag{2.7}$$

$k_{i,bending}$ and $\theta_{i,o}$ are the bending force constants and the natural angle of three atoms in an ideal strain free bonds, respectively. This equation is valid for ca. 10 degrees deformation. At larger deformation in a strained molecule a cubic term can correct the potential Eq. 2.8

$$V_{bending}=\sum_i \frac{1}{2}k_{i,bending}(\theta_i-\theta_{i,o})^2\left[1-k'_{i,bending}(\theta_i-\theta_{i,o})-k''_{i,bending}(\theta_i-\theta_{i,o})^2\ldots.\right] \tag{2.8}$$

$k_{i,bending}$, $k'_{i,bending}$, $k''_{i,bending}$ are the bending force constants.

Considering a rotation around a bond by a 0 to 180 degree, we can find minima and maxima (see the rotational energy profile of ethane, *Figure 2.3.*). In ethane molecule the staggered conformations are connected by eclipsed conformations.

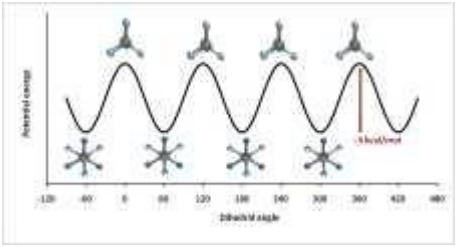


Figure 2.3. Rotational energy profile in ethane (Potential energy vs. dihedral angle of H-C-C-H)

The torsion energies can be described by a Fourier series of terms in Eq. 2.9

$$V_{torsion}=\sum_{n=0}^{N}\frac{V_n}{2}(1+\cos(n\omega-\gamma)) \tag{2.9}$$

ω is the dihedral angle, γ is the phase factor (torsion angle at minimum). n is the number of different rotational positions. At sp^3 carbon atoms (C-C-C-C) n = 3, $\gamma = 0^\circ$, at $sp^{2\cdot}$ (C-C=C-C), n = 2, $\gamma = 180^\circ$.

In AMBER FF three torsion expression is used

$$V_{torsion}=\frac{V_1}{2}(1+\cos\omega)+\frac{V_2}{2}(1+cos2\omega)+\frac{V_3}{2}(1+cos3\omega) \tag{2.10}$$

The three terms have different meanings: (i) the first expression is the description of the dipole-dipole, van der Waals and other interactions between atoms, (ii) the second expression is on the conjugation and/or hyperconjugation, (iii) the third expression is on the steric effects between 1,4 atoms.

Figure 2.4. Molecular mechanics handled by all-atom and united atom models

The molecules can be handled by all-atom and united atom model (see *Figure 2.4.*). It means that the models consist of the all-atoms in the molecules or only the heavy atoms (not H-atoms) and the H-atoms connected to polar atoms (not C-atoms, but N-, O-, S-, P-atoms with higher electronegativity than C-atoms). The charges, masses and force constants are corrected to the all-atom parameters. It is important to decrease the degrees of freedom for the geometry optimization (see Chapter 3).

The improper torsions and out-of–plane bending motion (see *Figure 2.5.*) is for stabilizing ring structures and the chirality of the structures (e.g. at the united atom model without out-of-plane restriction the chirality of the C_α can be changed).

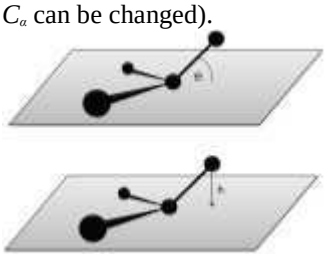


Figure 2.5. Out-of-plane bending with θ angle

The *Voop* potential can be handled by a quadratic equation Eq. 2.11.

$$V_{oop}=k_{oop}\theta^2 \tag{2.11}$$

The *Voop* potential is given by Eq. 2.12, where ω is the torsion angle starting in the centre.

$$V_{oop}=k_{oop}(1-cos2\omega) \tag{2.12}$$

In class 1, 2 and 3 force fields cross terms are applied: stretch-stretch, stretch-bend and bend-bend

$$V_{i1,i2}=\sum_{i1,i2}\frac{k_{i1,i2}}{2}(r_{i1}-r_{i1,0})(r_{i2}-r_{i2,0}) \tag{2.13}$$

$$V_{i1,i2,\theta}=\sum_{i1,i2}\frac{k_{i1,i2,\theta}}{2}(r_{i1}-r_{i1,0})(r_{i2}-r_{i2,o})(\theta-\theta_0) \tag{2.14}$$

In the Urey-Bradley force field the 1,3-nonbonded interactions are handled explicitly.

$$V^{UB}=\frac{1}{2}\sum_i K_i(\Delta r_i)^2+\frac{1}{2}\sum_i L_i(\Delta\alpha_i)^2+\frac{1}{2}\sum_i M_i(\Delta\rho_i)^2+\sum_i k_i\Delta r_i+\sum_i l_i\Delta\alpha_i+\sum_i m_i\Delta\rho_i \tag{2.15}$$

$\Delta\rho_i$ is the change in the distance between non-bonded atoms. K_i, L_i, M_i, k_i, l_i and m_i are the parameters of the force field.

2.1. Non-bonded interactions

The non-covalent interactions include the Coulomb interactions between point charges and the van der Waals interactions with 12-6 Lennard-Jones potential (at systems with a lot of H-bonds 10-6 Lennard-Jones potential). The functions are described in Eq. 2.16 and Eq. 2.17.

$$V_{Coulomb}=\sum_{i<j}^{N}\frac{q_iq_j}{4\pi\varepsilon r_{ij}} \tag{2.16}$$

r_{ij} is the distance between atoms i and j. q_i and q_j are the point charges of the atoms. ε is the effective dielectric constant.

$$V_{vdWaals}(r_{ij})=\varepsilon_{ij}\left[\left(\frac{R_{0,ij}}{r_{ij}}\right)^{12}-2\left(\frac{R_{0,ij}}{r_{ij}}\right)^{6}\right] \tag{2.17}$$

r_{ij} is the distance between atoms i and j. ε_{ij} is expressed by the arithmetic or geometric average of the van der Waals constants of the two atoms in the atom-pairs ($\varepsilon_{ij}=(\varepsilon_i\,\varepsilon_j)1/2$ or $\varepsilon_{ij}=(\varepsilon_i+\varepsilon_j)/2$), which depends ont he type of the force field). $R_{0,ij}$ is expressed similarly by deriving the parameters from the van der Waals parameters of *i* and *j* atoms (by the arithmetic or geometric average). The potential energy of the van der Waals interaction vs. distance of atoms is described in *Figure 2.6.* (12-6 Lennard-Jones potential, curve is black). In some cases r^{-9} is the repulsive function. There is a possibility to express the van der Waals interactions by a Buckingham exponential-r^{-6} potential Eq. 2.18.

$$V_{vdW}=\sum_{i,j}\left(c_1\exp(-c_3r_{ij})-c_2r_{ij}^{-6}\right) \tag{2.18}$$

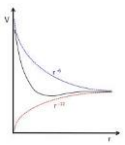


Figure 2.6. The potential energy vs. distance of atoms calculated by 12-6 Lennard-Jones function

There are many functional forms of the force fields. There are force fields for general organic compounds (MM2 [6], MM3 [7], MM4 [8], UFF (Universal Force Field) [9], MMFF [10].A lot of other force fields were developed for biomolecules (peptides, proteins, DNA, RNA): AMBER (AMBER94 [11], AMBER98 [12], AMBER99 [13], AMBER2002 [14]), CHARMM (CHARMM19 [15], CHARMM22 [16], GROMOS [17], OPLSAA, OPLSUA [18].

2.2. The MM force fields

The first generally used force field was developed by Allinger et al. [6-8]. A detailed description of the functions in MM2 and MM3 is summerized e.g. in Lit. [19]. The potential function is the same as Eq. 2.2. The electrostatic interactions are considered in a molecule the interactions between bond dipoles defined by Eq. 2.19 obtained by statistical mechanics.

$$V_{dipole}=\frac{\mu_1\mu_2}{\varepsilon r_{ij}^3}(\cos\zeta-3\cos\theta\cos\theta') \tag{2.19}$$

where ε is the effective dielectric constant in the solution. The angles are defined in *Figure 2.7*. The dipoles are modified by the electronegativity of the two heavy atoms. The MM2 force field is modified in MMX by Gilbert implemented in PCMODEL [20] which is useful for metal complexes, transition states, ions, too.

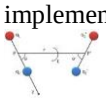


Figure 2.7. The interactions between two dipoles with the geometric parameters. The function is defined in Eq. 2.19.

2.3. The AMBER force field

Assisted Model Building with Energy Refinement (AMBER) based ont he following equation which is suitable for the biomolecules (peptides, proteins, DNA) [11-14]:

$$V(\mathbf{R})=\sum_{bonds}K_r(r-r_{eq})^2+\sum_{angle}K_\theta(\theta-\theta_{eq})^2+\sum_{dihedrals}\frac{V_n}{2}[1+\cos(n\phi-\gamma)]+\sum_{i<j}^{atoms}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+\sum_{i<j}^{atoms}\frac{q_iq_j}{\varepsilon r_{ij}} \tag{2.20}$$

Force fields based on atom-centered dipole polarizabilities can be applied by using the polarization term Eq. 2.21 polarization

$$V_{polarization}=-\frac{1}{2}\sum_{i}^{atom}\mu_i\cdot\mathbf{E}_i^{(0)} \qquad \text{polarization} \tag{2.21}$$

Where μ_i is induced atomic dipole, $\mathbf{E}_i$ is the the electric field.

A non-periodic simulation of aqueaus (implicit) solvation effect can be handled by means of the modification of the Coulomb interaction to Eq. 2.22:

$$V_{Coulomb}=\sum_{i<j}^{atom}\frac{q_iq_j}{f^{gb}(r_{ij})}+\sum_{i}^{atom}\sigma_iA_i \tag{2.22}$$

The first part is the responsible for the polar part of the solvation free energy, the second part is the non-polar contribution which depends on the atomic suface areas of the solvent accessible surface. σ_i is the atomic solvation free energy increment.

The force field is implemented in AMBER molecular mechanics/molecular dynamics package and in the package for preparation the input files (AMBERTOOLS) [21].

2.4. Charges

Charges int he traditional force fields were developed as point charges. In AMBER force field [11-14,21] the effective charges were obtained by fitting the gas phase electrostatic potential of small peptides calculated by HF/6-31G* and used RESP (Restrained Electrostatic Potential) or RESP-like charges were developed [22]. The charges means that how many electrons are shared between atoms. The calculation is not simple. The calculation is based on the following equation.

$$\phi_{esp}(\mathbf{r})=\phi_{nucl}(\mathbf{r})+\phi_{eec}(\mathbf{r})=\sum_{A}^{nuc}\frac{Z_A}{|\mathbf{R_A}-\mathbf{r}|}-\int\frac{d\mathbf{r}'\rho(\mathbf{r})}{|\mathbf{r}'-\mathbf{r}|}d\mathbf{r}' \tag{2.23}$$

A least square minimization of the potential at a point and the calculated potential with weighting factors for the points give the point charges. The Nth charge can be calculated

$$q_N = Z - \sum_{j=1}^{N-1} q_j \tag{2.24}$$

Another algorithm is also applied (Charges from Electrostatic Potentials using a Grid based method, CHELPG [23]).

The potential map can be calculated by ab initio method. Point charges with the same quality can be generated by semiempirical quantum chemical method AM1 with afitting function (AM1-BCC charges [24]).

2.5. Parametrization

The parametrization of the traditional force fields based on two approaches: (i) evaluation of experimental data and (ii) evaluation of theoretical calculations [19]. The parameters must be consistent in a force field. The experimental and theoretical data is fitted by the functioons applied int he force field.

2.6. Thermochemistry in Molecular Mechanics

It is very important to obtain the heat of formation of the molecules. The simplest calculation is related to alkanes. The heat of formation for an alkane can be given by Eq. 2.25.

$$\Delta H_f^0 = \Delta H_{sterics} + \Delta H_{conf} + \Delta H_{bond} + \Delta H_1^0 + \Delta H_2^0 + \Delta H_3^0 + \Delta H_4^0 \tag{2.25}$$

The ΔH_{steric} is calculated by molecular mechanics. ΔH_{conf}

$$\Delta H_{conf} = \sum_i N_i \Delta H_i \tag{2.26}$$

where N_i is the mole fraction of the conformers, ΔH_i is the enthalpy difference between conformers. ΔH_1^0, ΔH_2^0, ΔH_3^0 and ΔH_4^0 are the corrections for primary, secondary, tertiary and quaternary carbon atoms. ΔH_{bond} is the sum of the bond enthalpies for C-C and C-H bonds.

The heat of formation of alkanes is important, because the experimental values are very precise in calorimetric measurements. The molecules with sighly strained structure are good examples for the parametrization. The angle deformation and the torsion function can be refined on the basis of the experimental results [25]. The heat of formation of molecules with heteroatoms at one conformation is given in Eq. 2.27:

$$\Delta H_f^0 = \Delta H_{sterics} + \Delta H_{bond} \tag{2.27}$$

A similar expression without the steric contribution is available for the entropy estimation.

3. Non-Traditional (Polarizable) Molecular Mechanics Methods

The non-traditional molecular mechanics methods consider the polarizability of the atoms/groups int he molecules. We describe two of these methods: Ponder et al. [26] and Gresh et al. [27] developed AMOEBA [26] and SIBFA (Sum of Interactions between Fragments Ab Initio calculated) [27], respectively.

3.1. AMOEBA

The method AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) expand the classical force field expressions by the permanent electrostatic interaction between the point charges and multipoles and the induces electrostatic interactions considering polarizability (see Eq. 2.21). The expression contains the bond deformations, the bond-angle cross term, a formal Wilson-Decius-Cross decomposition of angle bending into in-plane (V_{angle}) and out-of-plane (V_{oop}) terms. The van der Waals expression is a Lennard-Jones potential modified to 14-7. The electrostatics decomposed into permenant and induced potential.

$$V = V_{bond} + V_{angle} + V_{bond-angle} + V_{oop} + V_{torsion} + V_{vdW} + V_{ele}^{perm} + V_{ele}^{ind} \tag{2.28}$$

The bond deformations are calculated by Eq. 2.27, which is a modified/refined Hooke's law with third and forth order polynom. bonds

$$V_{bond}=K_{bond}(r-r_o)^2\left[1-2.55(r-r_o)+\frac{7}{12}2.55(r-r_o)^2\right] \qquad \text{bonds} \tag{2.29}$$

$$V_{angle}=K_0(\theta-\theta_o)^2\left[1-0.014(\theta-\theta_o)+5.6\cdot10^{-5}(\theta-\theta_o)^2-7.0\cdot10^{-7}(\theta-\theta_o)^3+2.2\cdot10^{-8}(\theta-\theta_o)^4\right] \qquad \text{angle} \tag{2.30}$$

(2.22) out-of-plane

$$V_{torsion}=\sum_{torsion}\sum_{n}\frac{V_n}{2}[1+\cos(n\varphi-\gamma)] \qquad \text{torsion} \tag{2.31}$$

$$V_{oop}=0.02191418k_\chi\chi^2 \qquad \text{out-of-plane} \tag{2.32}$$

The van der Waals potential van der Waals

$$V_{vdW}=\varepsilon_{ij}\left(\frac{1.07}{\rho_{ij}+0.07}\right)^7\left(\frac{1.12}{\rho_{ij}+0.12}-2\right) \qquad \text{van der Waals} \tag{2.33}$$

The 14-7 function provide a softer repulsive character than the Lennard-Jones 6-12 function. It fits better the *ab initio* quantum chemical results and the liquid propertis in noble gases. In the expression all of the atom pairs are considered but the X-H bond distance is reduced (reduction factor) ont he basis of X-ray structural analysis.

$$\varepsilon_{ij}=\frac{4\varepsilon_{ii}\varepsilon_{jj}}{\left(\varepsilon_{ii}^{1/2}+\varepsilon_{jj}^{1/2}\right)} \qquad \text{and} \qquad R_{ij}^o=\frac{(R_{ii}^o)^3+(R_{jj}^o)^3}{(R_{ii}^o)^2+(R_{jj}^o)^2} \tag{2.34}$$

In the **permanent electrostatic interactions** the permanent multipoles (PAMs) are considered.by Eq. 2.35 which consist of point charges, dipole moment vectors and the quadrupoles.

$$M_i=\left[q_i,\mu_{ix},\mu_{iy},\mu_{iz},Q_{ixx},Q_{ixy},Q_{ixz},\ldots,Q_{izx},Q_{izy},Q_{izz}\right]^T \tag{2.35}$$

$\partial q_i/\partial x_i$, $\partial q_i/\partial y_i$, $\partial q_i/\partial z_i$ are the dipole moments related to the Descartes coordinates, the second derivatives are the quadrupole moments as described in Eq. 2.36.

$$T_{ij}=\begin{bmatrix} 1 & \frac{\partial}{\partial x_i} & \frac{\partial}{\partial y_i} & \frac{\partial}{\partial z_i} & L \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial z_i\partial x_j} & \frac{\partial^2}{\partial z_i\partial y_j} & \frac{\partial^2}{\partial z_i\partial z_j} & L \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i\partial x_j} & \frac{\partial^2}{\partial y_i\partial y_j} & \frac{\partial^2}{\partial y_i\partial z_j} & L \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i\partial x_j} & \frac{\partial^2}{\partial z_i\partial y_j} & \frac{\partial^2}{\partial z_i\partial z_j} & L \\ M & M & M & M & O \end{bmatrix}\frac{1}{r_{ji}} \tag{2.36}$$

In Cartesian polytensor formalism, the interaction energy between atoms *i* and *j* with r_{ij} distance between them is $V_{perm}^{elec}(r_{ij})=M_i^TT_{ij}M_j$. T_{ij} is the tensor defined by Eq. 2.36. Atomic multipole moments are derived by using Stone's Distributed Multipole Analysis (DMA) [28].

3.2. SIBFA

Inter and intramolecular interaction energy (ΔE)

$$\Delta E=E_{MTP}+E_{rep}+E_{pol}+E_{ct}+E_{disp} \tag{2.37}$$

where E_{MTP} denotes the multipolar electrostatic energy contribution, E_{rep} is the short range repulsion energy calculated for bond-bond, bond-lone pair and lone pair-lone pair interactions, E_{pol} is the polarization energy contribution calculated by the distributed, anisotropic polarizabilities on the constitutive fragments [27]. The polatizabilities are distributed on the localized orbitals using the method of Garmer and Stevens [28]. $E_{ct,}$ is the charge transfer energy contribution and E_{disp} is the dispersion energy contribution. The parameters are

summerized in a library which is based on *ab initio* calculations. The biomolecules are in solution, which means that the system must be in solution. The correction is the solvation energy:

$$E_{solv} = E_{cav} + E_{el} + E_{pol} + E_{dr} \quad (2.38)$$

E_{cav} is the cavitation energy, E_{el} is the solvent-solute electrostatic energy, E_{pol} is the solute polarization energy, E_{dr} is the dispersion-repulsion energy contribution energy.

SIBFA and SIBFA/Continuum method [25] are the simulated *ab initio* calculations for amino acids in different solvents. The results are excellent for small peptides, DNA and their Zn^{2+} and Cu^{2+} ion complexes.

The non-traditional MM/MD methods demand more CPU time than the traditional methods. Though the traditional methods are not very precise, their application for a long simulation describe more precision than that of the non-traditional methods.

4. Summary

The chapter dealt with a basic method to learn the structure of molecules, intra- and intermolecular interactions which can modify the expected structures. A simple and fast method for the evaluation of the structure and the basic properties of large molecules and macromolecular systems (protein-protein, protein-DNA, protein-ligand associations, etc.)

5. References

1. Allinger NL, Burkert U (1982). Molecular Mechanics. An American Chemical Society Publication. ISBN 0-8412-0885-9.

2. A. T. Hagler , P. S. Stern , S. Lifson , S. Ariel, Urey-Bradley force field, valence force field, and ab initio study of intramolecular forces in tri-tert-butylmethane and isobutane. J. Am. Chem. Soc., 101 (4), 813–819(1979).

3. Schlick T (2002). Molecular modeling and simulation: an interdisciplinary guide. Berlin: Springer. ISBN 0-387-95404-X.

4. C.D. Sherill, Introduction to Molecular Mechanics, School of Chemistry and Biochemistry, Georgia Institute of Technology

5. J. J. Gajewski, K. E. Gilbert, J. McKelvey, MMX an enhanced version of MM2, Adv. in Molecular Modeling, A Research Annual. Ed. by D. Liotta, Vol. 2, 1990. JAI Press Inc.

6. a) N. L. Allinger, Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms, J. Am. Chem. Soc., 99, 8127-8134 (1977). b) J. T. Sprague, J. C. Tai, Y. Yuh and N. L. Allinger, The MMP2 Calculational Method, J. Comput. Chem., 8, 581-603 (1987). c) N. L. Allinger, R. A. Kok and M. R. Imam, Hydrogen Bonding in MM2, J. Comput. Chem., 9, 591-595 (1988).

7. a) N. L. Allinger, Y. H. Yuh and J.-H. Lii, Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1, J. Am. Chem. Soc., 111, 8551-8566 (1989). b) J.-H. Lii and N. L. Allinger, Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 2. Vibrational Frequencies and Thermodynamics, J. Am. Chem. Soc., 111, 8566-8575 (1989). c) J.-H. Lii and N. L. Allinger, Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 3. The van der Waals' Potentials and Crystal Data for Aliphatic and Aromatic Hydrocarbons, J. Am. Chem. Soc., 111, 8576-8582 (1989). d) N. L. Allinger, H. J. Geise, W. Pyckhout, L. A. Paquette and J. C. Gallucci, Structures of Norbornane and Dodecahedrane by Molecular Mechanics Calculations (MM3), X-ray Crystallography, and Electron Diffraction, J. Am. Chem. Soc., 111, 1106-1114 (1989). e) N. L. Allinger, F. Li and L. Yan, Molecular Mechanics. The MM3 Force Field for Alkenes, J. Comput. Chem., 11, 848-867 (1990). f) N. L. Allinger, F. Li, L. Yan and J. C. Tai, Molecular Mechanics (MM3) Calculations on Conjugated Hydrocarbons, J. Comput. Chem., 11, 868-895 (1990). g) J.-H. Lii and N. L. Allinger, Directional Hydrogen Bonding in the MM3 Force Field. I, J. Phys. Org. Chem., 7, 591-609 (1994). h) J.-H. Lii and N. L. Allinger, Directional Hydrogen Bonding in the MM3 Force Field. II, J. Comput. Chem., 19, 1001-1016 (1998).

8. Norman L. Allinger , Kuohsiang Chen, Jenn-Huei Lii, An improved force field (MM4) for saturated hydrocarbons. J. Comput. Chem. 642–668(1996).

9. A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard III, W.M. Skiff, UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. J.Am. Chem. Soc. 114 10024–10035(1992).

10. a) T. A. Halgren, Merck Molecular Force Field. I. Basis, Form, Scope,Parametrization, and Performance of MMFF94, J. Comput. Chem., 17, 490-519 (1995). b) T. A. Halgren, Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions, J. Comput. Chem. 17, 520-552 (1995). c) T. A. Halgren, Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94, J. Comput. Chem. 17, 553-586 (1995). d) T. A. Halgren and R. B. Nachbar, Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94, J. Comput. Chem., 17, 587-615 (1995). e) T. A. Halgren, Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules, J. Comput. Chem. 17, 616-641 (1995).

11. a) W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, J. Am. Chem. Soc., 117, 5179-5197 (1995). b) G. Moyna, H. J. Williams, R. J. Nachman and A. I. Scott, Conformation in Solution and Dynamics of a Structurally Constrained Linear Insect Kinin Pentapeptide Analogue. Biopolymers, 49, 403-413 (1999). c) W. S. Ross and C. C. Hardin, Ion-Induced Stabilization of the G-DNA Quadruplex: Free Energy Perturbation Studies. J. Am. Chem. Soc., 116, 6070-6080 (1994). d) Current parameter values are available from the Amber site, located at http://ambermd.org/

12. a) T. E. Cheatham III, P. Cieplak and P. A. Kollman, A Modified Version of the Cornell et al. Force Field with Improved Sugar Pucker Phases and Helical Repeat, J. Biomol. Struct. Dyn., 16, 845-862 (1999). b) W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman,A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, J. Am. Chem. Soc., 117, 5179-5197 (1995). c) G. Moyna, H. J. Williams, R. J. Nachman and A. I. Scott, Conformation in Solution and Dynamics of a Structurally Constrained Linear Insect Kinin Pentapeptide Analogue. Biopolymers, 49, 403-413 (1999).

13. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling,Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters, PROTEINS, 65, 712-725 (2006).

14. P. Cieplak, J. Caldwell and P. Kollman, Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases, J. Comput. Chem., 22, 1048-1057 (2001).

15. a) W. E. Reiher III, Theoretical Studies of Hydrogen Bonding, Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge, MA, 1985. b) L. Nilsson and M. Karplus, Empirical Energy Functions for Energy Minimizations and Dynamics of Nucleic Acids, J. Comput. Chem., 7, 591-616 (1986). c) E. Neria, S. Fischer and M. Karplus, Simulation of Activation Free Energies in Molecular Systems., J. Chem. Phys., 105, 1902-1921 (1996).

16. a) A. D. MacKerrell, Jr., et al., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins, J. Phys. Chem. B, 102, 3586-3616 (1998). b) N. Foloppe and A. D. MacKerell, Jr., All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data, J. Comput. Chem., 21, 86-104 (2000).

17. a) W. F. van Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hünenberger, P. Krüger, P. A. E. Mark, W. R. P Scott, I. G. Tironi, Biomolecular Simulation: The GROMOS96 Manual and User Guide; vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: Zürich, Groningen, 1996. b) W.R.P. Scott, P.H. Huenenberger, I.G. Tironi, A.E. Mark, S.R. Billeter, J. Fennen, A.E. Torda, T. Huber, P. Krueger and W.F. van Gunsteren. The GROMOS Biomolecular Simulation Program Package, J. Phys. Chem. A, 103,3596-3607(1996).

18. a) W. L. Jorgensen and J. Tirado-Rives, The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin, J. Am. Chem. Soc., 110, 1657-1666 (1988). b) D. S. Maxwell, J. Tirado-Rives and W. L. Jorgensen, A Comprehensive Study of the Rotational Energy Profiles of Organic Systems by Ab Initio MO Theory, Forming a Basis for Peptide Torsional Parameters, J. Comput. Chem, 16, 984-1010 (1995) c) W. L. Jorgensen and D. L. Severance, Aromatic-Aromatic Interactions: Free Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene, J. Am. Chem. Soc., 112, 4768-4774 (1990) d) S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr. and P. Weiner, A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins, J. Am. Chem. Soc., 106, 765-784 (1984) e) S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case, An All Atom Force Field for Simulations of Proteins and Nucleic Acids, J. Comput. Chem., 7, 230-252 (1986). f) L. X. Dang and B. M. Pettitt, Simple Intramolecular Model Potentials for Water, J. Phys. Chem., 91, 3349-3354 (1987). g) W. L. Jorgensen, J. D. Madura and C. J. Swenson, Optimized Intermolecular Potential Functions for Liquid Hydrocarbons, J. Am. Chem. Soc., 106, 6638-6646 (1984). h) E. M. Duffy, P. J. Kowalczyk and W. L. Jorgensen, Do Denaturants Interact with Aromatic Hydrocarbons in Water? J. Am. Chem. Soc., 115, 9271-9275, (1993). i) W. L. Jorgensen, C. J. Swenson, Optimized Intermolecular Potential Functions for Amides and Peptides. Structure and Properties of Liquid Amides, J. Am. Chem. Soc., 107, 569-578 (1985).

19. J. P. Bowen, N. L. Allinger, Molecular Mechanics: The Art and Science of Parametrization, pp. 81-98, in Reviews in Computational Chemistry II, Ed. by K. B. Lipkowitz, D. B. Boyd, VCH, 2007.

20. PCMODEL, Serena Software, Bloomington, USA, 2008.

21. D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman (2012), AMBERTOOLS 12 and AMBER 12, University of California, San Francisco.

22. a) E. Vanquelef, S. Simon, G. Marquant, J.C. Delepine, P. Cieplak and F.-Y. Dupradeau, R.E.D. Server: a web service designed to automatically derive RESP and ESP charges and to generate force field libraries for new molecules and molecular fragments, Université de Picardie Jules Verne - Sanford-Burnham Institute of Medical Research, 2009, http://q4md-forcefieldtools.org/REDS b) C. Cézard, E. Vanquelef, P. Cieplak and F.-Y. Dupradeau, Tutorials describing the use of the Ante_RED.-1.x and R.E.D. III.x programs, the R.E.DD.B database and R.E.D. Server, Université de Picardie Jules Verne - Sanford-Burnham Institute of Medical Research, 2007, http://q4md-forcefieldtools.org/Tutorial. c) J. Wang, P. Cieplak and P. A. Kollman, How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calcluating Conformational Energies of Organic and Biological Molecules?, J. Comput. Chem., 21, 1049-1074 (2000). d) W. D. Cornell, P. Cieplak, C. Bayly, P. A. Kollmann, Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation, J. Am. Chem. Soc., 115 (21), 9620–9631(1993). c) A. Laio, J. VandeVondale, U. Rothlisberger, D-RESP: Dynamically Generated Electrostatic Potential Derived Charges from Quantum Mechanics/Molecular Mechanic Simulations, J. Phys. Chem. B, 106, 7300-7307(2002). d) R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments

23. C. M. Breneman, K. B. Wiberg, Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J. of Comp. Chem. 11 (3), 361-383 (1994).

24. A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method, J. Comp. Chem. 21(2), 132-146(2000).

25. Donald W. Rogers, Molecular Mechanics in Computational Thermochemistry, Computational Thermochemistry, ACS Symposium Series, Vol. 677. Chapter 7, pp 119–140, 1998.

26. a) P. Ren, C. Wu and J. W. Ponder, Polarizable Atomic Multipole-based Potential for Proteins: Model and Parameterization, in preparation. b) P. Ren, C. Wu and J. W. Ponder, Polarizable Atomic Multipole-based Potentials for Organic Molecules, in preparation c) J. W. Ponder and D. A. Case, Force Fields for Protein Simulation, Adv. Prot. Chem., 66, 27-85 (2003). d) P. Ren and J. W. Ponder, Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation, J. Phys. Chem. B, 107, 5933-5947 (2003). e)

P. Ren and J. W. Ponder, A Consistent Treatment of Inter- and Intramolecular Polarization in Molecular Mechanics Calculations, J. Comput. Chem., 23, 1497-1506 (2002). f) J. Wang, P. Cieplak and P. A. Kollman, How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calcluating Conformational Energies of Organic and Biological Molecules? J. Comput. Chem., 21,1049-1074 (2000).

27. a) N. Gresh, B.-P. Roques, Thermolysin-Inhibitor Binding: Effect of the His230-> Ala Mutation ont he Relative Affinities of Thiolate Versus Phosphoramidate Inhibitors – A Model Theoretical Investigation Incorporating Continuum Reaction Field Hydration Model. Biopolymers, 41, 145-164 (1977). b) N. Gresh, Inter- and intramolecular interactions. Inception and refinements of the SIBFA, molecular mechanics (SMM) procedurem a separable, polarizable methodology grounded on ab initio SCF/MP2 computations. Examples of applications to molecular recognition problems. J. Chim. Phys., 94, 1365-1416(1997).

28. A. J. Stone, Distributed Multipole Analysis, or How to Describe a Molecular Charge Distribution, Chem. Phys. Letters , 83, 233-239(1981).

29. D.R. Garmer, W. J. Stevens, Transferability of molecular distributed polarizabilities from a simple localized orbital based method, J. Phys. Chem. 93(25), 8263-8270(1989).

6. Further Readings

1. A. K. Rappé, C. J. Casewit, Molecular Mechanics across Chemistry, University Science Book, CA, USA, 1997. ISBN0-935702-77-6.

2. D. A. Case, T. A. Darden, T. E. Cheathem III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, B. Wang, D. A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross and P. A. Kollman, Amber8. User's Manual, University of California, San Francisco, 2004.

3. C. J. Cramer, Essentials of Computational Chemistry, Theories and Models", John Wiley and Sons, LTD, ISBN 0-471-48552-7. Chapter 2, Chapter 3.

4. D. M. Hirst, A Computational Approach to Chemistry, Blackwell Scientific Publications, Oxford, London, ISBN 0-632-02433-6. 1990. Chapter 3.

5. G. H. Grant, W. G. Richards, Computational Chemistry, Oxford Science Publications, Ocford Chemistry Primers, Oxford University Press, 1995. Chapter 3.

6. G. M. Keserű, I. Kolossváry, Molecular Mechanics and Conformational Analysis in Drug Design, Blackwell Science, Oxford, ISBN0632052899, 1999.

7. C. L. Brooks, M. Karplus, B. M. Petitt, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, Series Editors I. Prigogine and S. A. Rice, Advances in Chemical Physics, John Wiley&Sons, 1987.

7. Questions

1. What kind of deformations can strain in a molecule?

2. Which function describe the small covalent deformations?

3. Please, compare the bond deformations of two bonds graphically: (i) $k_{C=C}$ = 40.2 kJ/(Å^2 mol), $l_{0,C=C}$ = 1.337 Å, (ii) k_{H-C} = 19.3 kJ/(Å^2 mol), $l_{0,H-=C}$ = 1.090 Å.

4. Which pair potentials describe the non-covalent interactions?

5. Does the pair potentials give the real interaction energies?

6. What is the steric energy?

7. What is the main difference between the traditional and non-traditional force fields? What energy partitioning are used int he two cases?

8. What is the same and differences in SIBFA and in AMOEBA?

8. Glossary

Hooke's law: It describe the deformations in bonds and angles by a harmonic oscillator equation.

Force field: The force field includes the functions and the parameters of the functions to describe the deformation of the molecules.

Bond deformations: The bond deformations are described as a harmonic function by the Hook's law.

Angle deformation: The angle deformations are described as a harmonic function.

Non-bonded interactions: Electrostatic (multipole) interactions and van der Waals interactions. Coulomb function or multipole interaction functions are used int he previous, 12-6, 10-6 Lennard-Jones or Buckingham potential is applied for the calculations.

Point charges: Most of the traditional force fields apply point charges. Atoms are not points, that is why RESP or CHELP method is used to estimate the point charges.

Polarizability: Most of the atoms have polarizability. On the effect of charges induced charges can be formed.

Multipole interactions: Interactions of point-charges-point charges, dipole-point charges, dipole-dipole, point charges-quadrupole, dipole-quadrupole, etc. interaction are multipole interactions.

Chapter 3. Electrostatics in Molecules

(Tamás Körtvélyesi)

Keywords: molecular electrostatics, non-linear Poisson-Boltzmann equation (NLPBE), linear Poisson-Boltzmann equation (LPBE), Tanford-Kirkwood equation (TKE), numerical solution, molecular surface

What is described here? This chapter includes the description of the electrostatic interactions between charged groups with the extended extrapolation of size scales

What is it used for?The prediction of the electrostatic properties of these large molecules (e.g. peptides, proteins, DNAs, PNAs, etc.) can help in modelling the association of these molecules (electrostatic complementarity) and the solvation free energy (implicit solvation models). The calculation of the difference between the bound and unbound states with and without native or drug-like ligands and the configurational properties in solutions are very important if we do not consider the explicit solvents (see Chapter 4). The developement of the the fast computational methods with different approaches for the solution of the Poisson-Boltzmann equation (PBE) and the Tanford-Kirkwood equuation (TKE) are important for the knowledge of the interaction (association) energies. The method is appropriate to simulate the motion of the biomolecules (diffusion) by means of molecular dynamics (MD).

What is needed? To elucidate this chapter the knowledge of the biologically important molecules and their intramolecular interactions (Chapter 1), molecular mechanics (Chapter 2) are necessarry. Also, the basic knowledge of the solution of the differential equations, the basic physics, physical chemistry and organic chemistry is important. In the end of this chapter, the basic concepts of the calculations of the electrostatics in large molecules will be attained which is the basis in the the prediction of solvation free energy (Chapter 4) and the association free energy of (protein-protein, protein-ligand, protein-DNA, protein-PNA, etc.) molecules (Chapter 9)

1. Introduction

The atoms in molecules have partial charges, which direct the association of these molecules and the solvation of the molecules in different solvents. The counter ions (ionic strength) in the solution have also effect on the interactions. The solution of PBE and/or TKE make possible to obtain the solvation free energy, the association energy with ligands and the association of peptides/proteins with metal (gold, silver, etc.) surfaces in the nano scale with modeling of the field of solvents.The method is simpler than the all atom models with explicit water molecules (see Chapter 4), but it does not contain some real effects in the solution (e.g. viscosity, the real interactions between the solvent molecules and the solute, etc.). The method is suitable for calculating the potential around an extended structure, too. The solution includes grid calculations with fixed grid space, grid space for refinement or multigrid space.

2. Coulomb Equation

The electrostatic pair potential ($\varphi_i(\mathrm{r})$) around a point charge (q_i) in a homogeneous medium with an ε

effective dielectric constant ($\varepsilon = \varepsilon_r \varepsilon_o$,where ε_r is the relative dielectric constant and ε_o is the dielectric constant *in vacuo*) is

$$\varphi_{i_i}(\mathrm{r}) = \frac{1}{4\pi\varepsilon}\frac{q_i}{|\mathrm{r}-\mathrm{r}_i|} \tag{3.1}$$

where ε_o = 8.854 x 10^{-12} C^2 N^{-1} m^{-2}. The location of the charge *i* is r_i. The sign is negative at negative charge and positive at positive charge. The electrostatic potential in a system with N point charges

$$\varphi_i(\mathrm{r}) = \frac{1}{4\pi\varepsilon}\sum_i^N \frac{q_i}{|\mathrm{r}-\mathrm{r}_i|} \tag{3.2}$$

The total electrostatic interactions in a protein [1] used the pairwise Coulomb's law in a system with homogeneous medium consisting of N point charges can be written as

$$\Delta G_{el}=\frac{1389.8\,kJ/mol}{\varepsilon_r}\sum_{i=1}^{N}\sum_{j<i}^{N}\frac{q_i q_j}{r_{i,j}} \tag{3.3}$$

ΔG_{el} is the electrostatic interaction energy (free energy) at room temperature relative to the energy between the point charges at infinite distances, ε_r is the relative dielectric constant. The sign is negative at the interaction of different charges (attractive) and positive at the same point charges (repulsive interactions) [2]. The interaction energies calculated by the Coulomb's law are pairwise interaction energies without considering the many-body interactions (see e.g. Axelrod-Teller's formula [3]). These formulas are not totally valid for extended structures with charges..

3. Poisson Equation

The electrostatic potential φ(r) in vacuo for an extended structure with arbitrary shape can be given by the Poisson Eq. 3.4.

$$\nabla^2\varphi(\mathbf{r})+4\pi\rho(\mathbf{r})/\varepsilon_o=0 \tag{3.4}$$

where ρ(r) is the charge density as a function of position (r is the Cartesian coordinates of a point in space, ε_o is the effective dielectric constant *in vacuo*. Nabla is defined by Eq. 3.5 as a vector operator

$$\nabla=\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right) \tag{3.5}$$

and

$$\nabla^2=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right) \tag{3.6}$$

In spherical coordinates the Laplacian operator is given by Eq. 3.7

$$\nabla^2=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} \tag{3.7}$$

and in cylindrical coordinates the operator is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}+\frac{\partial^2}{\partial z^2} \tag{3.8}$$

The electrostatic potential φ(r) in a uniform dielectric medium with a relative dielectric constant ε_r ($\varepsilon = \varepsilon_o\,\varepsilon_r$)

$$\nabla^2\varphi(\mathbf{r})+4\pi\rho(\mathbf{r})/\varepsilon=0 \tag{3.9}$$

where ρ(r) is the charge density and $\varepsilon = \varepsilon_o\,\varepsilon_r$ is the dielectric constant in a uniform media. The electrostatic potential φ(r) is

$$\varphi(r)=\iiint\frac{\rho(r)}{\varepsilon|\mathbf{r}-\mathbf{r}'|}dr \tag{3.10}$$

The integral is over the space.

The electric field is given by Eq. 3.11 [4]

$$\mathbf{E}=-\nabla\varphi \tag{3.11}$$

The Poisson equation becomes in a uniform media

$$\nabla\mathbf{E}=\frac{\rho}{\varepsilon\varepsilon_o} \tag{3.12}$$

If the effective dielectric constant, ε depends on the coordinates (ε(r) in a non-uniform media), Eq. 3.9 equation is modified to Eq. 3.13 as the general form of the Poisson equation.:

$$\nabla[\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})]+4\pi\rho(\mathbf{r})=0 \tag{3.13}$$

The solute of the equation is with a low relative dielectric constant of ε_r of about 2 to 4 (e.g. in protein, where protein can be considered as a dielectric) Eq. 3.13 can be used. In some cases ε_r was suggested to be of about 20. We can not say anything exactly on ε_r, the structure of proteins are different and that is why ε_r, can be also different. These values are only approximations. In a water solution ε_r is *ca*. 80 [5].

The relative dielectric constants are handled in water (protein in water) as it can be seen in *Figure 3.1*. The analytical solution is available only for spherical, cylindrical or planar structures . The spherical model of an ion can be seen in *Figure 3.2*. The relative dielectric constant of the media is ε_r.

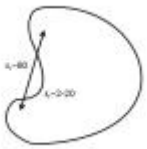


Figure 3.1. A shematic figure on the protein in water with the ε_r in the solvent and in the solute.

The spherical cylindtical model of an ion can be seen in *Figure 3.2*.

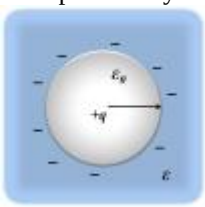


Figure 3.2. The spherical model of the ions.

The analytical solution of the model for sphericaln cylindrical, or planar symmetry is given in Eq. 3.14 - 3.15.

$$\varphi(\mathbf{r})=\left(\frac{q}{4\pi\varepsilon_o r}\right)-q\left(\frac{1}{\varepsilon_o}-\frac{1}{\varepsilon}\right)\frac{1}{4\pi a}, \qquad r<a \tag{3.14}$$

$$\varphi(\mathbf{r})=\left(\frac{q}{4\pi\varepsilon r}\right), \qquad r\geq a \tag{3.15}$$

The random positions of the ions around the large molecules can be described by the Boltzmann distribution.

4. Boltzmann Distribution

The charge density (ρ_{ion}(r)) is the sum of the charges in solutes and the mobile ions (Na^+, Cl^-, K^+, Ca^{2+}, etc.) in the solvent. The mobile ions in the solvent are handled to be uniform and the the Boltzmann distribution is used for the ion distribution:

$$\rho_{ion}(\mathbf{r})=e\sum_{i=1}^{m}c_i z_i \exp(-z_i e\varphi(\mathbf{r})/k_B T) \tag{3.16}$$

where e is the charge of the electron, z_i and c_i are the charge number and the concentration of the ions i in the bulk solution, respectively [1,4]. k_B is the Boltzmann constant. m is the number of the mobile ion species in the solution. The charge density of the solute can be determined as charges with fixed positions („source charges") (see Lit.[5]):

$$\rho_{\text{fixed}}(\mathbf{r})=e\sum_{i=1}^{M}q_i\delta(\mathbf{r}-\mathbf{r}_i) \tag{3.17}$$

δ(x) is the delta function. δ(x-y)=0 if x≠y and δ(x-y)=1 if x=y. M is the number of charges on biomolecules. In a one-to-one electrolyte (one positive and one negative ion, e.g. Na^+ and Cl^- ions) the Eq. 3.17 is simplified to

$$\rho_{ion}(r)=-\frac{\kappa'^2(r)}{4\pi}\sinh\left(\frac{e\varphi(r)}{k_B T}\right) \tag{3.18}$$

κ'(r) is the modified Debye-Hückel parameter is defined by Eq. 3.19. κ is the Debye-Hückel inverse length.

$$\kappa^2=\frac{\kappa'^2}{c(r)}=\frac{8\pi N_A e^2 I}{1000 k_B T} \tag{3.19}$$

where N_A is the Avogadro number, e is the electric charge, k_B is the Boltzmann constant, T is the temperature, I is the ionic strength of the bulk solution

$$I=\frac{1}{2}\sum_i c_i z_i^2 \tag{3.20}$$

c(r) is the concentration of the ions in the molecules with the fixed charges [6]. c_i and z_i are the concentration and the charge of the ions in the bulky solution, respectively. These expressions make possible to find the dependence between electrostatic potential (φ(r)) and the the charge density (ρ(r)).

5. Poisson-Boltzmann Equation (PBE)

The nonlinear form of the PBE [5-10] (NLBE) is a second order nonlinear elliptic partial differential equation . Analytical solution is available only for spheres and cylinders. For biomolecules, as proteins, DNA only numerical solution is possible.

The charge density („source charges") of the solute is given by the Eq. 3.21.

$$\left(\nabla[\varepsilon(r)\nabla\varphi(r)]-c(r)\kappa^2(r)\sinh[(\varphi(r))e/k_B T]+4\pi e\sum_{i=1}^{M} q_i\delta(r-r_i)=0\right) \tag{3.21}$$

It means that the electrostatic interactions between charges in biological systems depends on the ionic strength and the pH of the medium, too (see in Chapter 6). The first theoretical studies were published almost a century ago.

The analytic solution of Eq. 3.21. is available only for simple geometric objects. For complex systems it is possible to solve it by iterative finite difference methods (see later).

5.1. Linearized Poisson-Boltzman Equation (LPBE)

$$\sinh(x)=x+x^3/3!+x^5/5!+x^7/7!+\ldots \tag{3.22}$$

Considering sinh φ(r) ~ φ(r) (see Eq. 3.21), the linearized form of PBE (LPBE) can be obtained [11],

$$\left(\nabla[\varepsilon(r)\nabla\varphi(r)]-c(r)\kappa^2(r)(\varphi(r))e/k_B T+4\pi e\sum_{i=1}^{M} q_i\delta(r-r_i)=0\right) \tag{3.23}$$

Both the PBE and LPBE equations are determined by ε_r(r), c(r) and the positions of the atoms in molecules (q) [5]. The model of an ion pair can be seen on *Figure 3.3*.

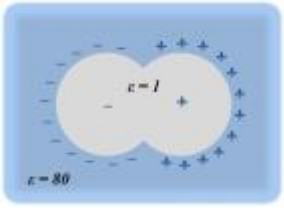


Figure 3.3. The model of the electrostatic calculation in an extended structure (e.g. in a protein)

One model can be described [6]:

Region 1: Inside the molecule. $\varepsilon_r(r)$ is 2 to 20., $c(r)$ is set to 0 (no ions) and q_i positioned by the atomic coordinates.

Region 2: Stern region. $\varepsilon_r(r)$ is set to the value in bulky solvent ($\varepsilon_r = 80$). It is supposed it is a region without ions. $c(r)$ is 0 (no ions). The source charge density is zero.

Region 3: *In bulk solvent*. In the bulk solvent the relative dielectric constant is 80. c is 1. The source charge density is zero.

The classical PBE does not include the possible difference in the size of ions. A modified PBE was developed which considers this difference („size modified PB (SMPB) equation") [12]. The orientation and strong dipolar moments of water molecules is described by the „dipolar Poisson-Boltzmann (DPB)" model [13]. The hydration forces, ionic associations and short range hydrophobic effects are calculated by the combination of SMPB and DPB methods [6].

5.2. Tanford-Kirkwood Equation (TKE)

The Tanford-Kirkwood equation [17-20] is a separeted partial differential equiation of PBE in two media. A model for proteins with the ε_1 relative dielectric constant in the molecule and ε_2 in the bulky solvent can be seen in *Figure 3.4.*

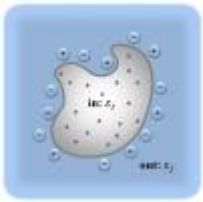


Figure 3.4. The model of the proteins in solution with counter ions

The TK equation can be seen in Eq. 3.27 and 3_28.

$$\nabla^2\varphi_1(r)=-\sum_{i=1}^{N}\frac{q_i}{\varepsilon_i}\delta(r-r_i) \tag{3.24}$$

$$\nabla^2\varphi_2(r)-\kappa^2\varphi_2(r)=0 \tag{3.25}$$

κ^2 is given by Eq. 3.19. $\kappa^2 \sim \beta I$.

6. Molecular Surface and Volume

In the calculation the surface and/or volume of the molecule is necessarry. The different surface and volume types are depicted on *Figure 3.5*. The radius of the probe sphere is 1 to 1.4 Å in water as solvent. The difference between the van der Waals surface, molecular surface and the solvent accessible surface (SAS) can be seen. The no-entrant surface is the part of the surface where the probe sphere does not reach the atoms. The fast calculation of the surface and volume is basic in the calculations. The grid calculation and the potential on the (van der Waals) surface is calculated by sophisticated methods [1]. The grid geometry and the interpolations of the potentials between the grid points (2D or 3D grids – trigonal, tetrahedral, tetraeder, …) to smooth the difference in dielectric constants) are also very important (see e.g. Lit. [6]). The grid spacing for calculating solvation energy is 0.2-0.3 Å in UHBD. Int he active centre the grid spacing can be refined.

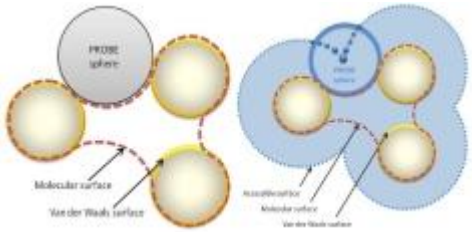


Figure 3.5. The molecular surface, van der Waals surface and the surface accessible surface (SAS). The probe sphere is the model of solvent.

7. Numerical Solution of non-linear Poisson-Boltzmann Equation (NPBE), Linear Poisson-Boltzmann Equation (LPBE) and Tanford-Kirkwood Equation (TKE)

The application of the results of PBE and LPBE is important to know (i) the electrostatic potential ont he surface of a biomolecule, (ii) the electrostatic potential outside the molecule, (iii) calculation of the free energy of a biomolecule and (iv) calculation of the electrostatic field to give the mean forces [18].

The analytical solution of PBE for real molecules as proteins, DNA, PNA, etc. are not available. Only numerical methods can give solutions. Several program packages for the solution of PBE were developed: DELPHI [19], UHBD (University of Houston Brownian Dynamics) [20], APBS (Adaptive Poisson-Boltzmann Solver) [21], MEAD [22], ZAP [23]. Two main methods are developed: (i) surface based methods and (ii) volume based methods [5].

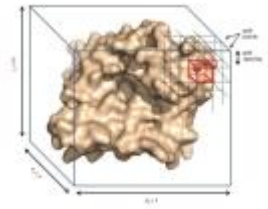


Figure 3.6. Grids for the solution of PBE with the +1 charge probe (PDB Id.: 1yet without ligand: geldanomycin and without structural water)

i. Solution of PBE on a surface mesh (**BEM**). The molecules („interior") and the space around the molecule is handled separately. The electrostatics of the interior is solved by the Poisson equation. The outside part is solved by the PB equation. The molecular surface is by the method of polygonal approach, triangular mesh (the discretization is in 2D). The interface between the two regions is handled by a continuous displacement field. The **BEM** method is faster than the volumetric methods. That is why the two methods are combined.

ii. **FD** and **FEM** approaches in the volumetric mesh. The discretization is in 3D space.

The application of the results of PBE and LPBE is important to know (i) the electrostatic potential on the surface of a biomolecule, (ii) the electrostatic potential outside the molecule, (iii) calculation of the free energy of a biomolecule and (iv) calculation of the electrostatic field to give the potential energy mean forces (PMF) [18], which describes how the free energy changes along a coordinate.

The electrostatics of the protein in the solvent is calculated by an iterative finite-difference approach with mapping on a cubic lattice with parameters $\rho(r)$, $\kappa(r)$ and $\varepsilon(r)$.

DelPhi [19]: With the Cartesian coordinates it calculates the electrostatic potential from the known geometry and the known charge distribution by using finite difference method in the solution of LPBE and full NPBE for proteins with arbitrary shape and charge considering the ionic strength. It considers ionic strength of the media. It can be used for extremly highly dimensions.

UHBD [20]: The electrostatic interactions can be calculated by LPBE or the full NPBE. The potential of the mean force is approximated by the electrostatic energy.with neclecting the non-electrostatics and the effect of ions. The most active part of the interactions can be refined by increasing the resolution of the meshes.

APBS [21]: The method is for the evaluation of the electrostatic properties for a wide range of length scale (tens to millions atoms).

MEAD [22]: The method solves the PBE and optionally calculates Brownian dynamics pKa of protein sidechains.

ZAP [23]: The algorithm calculates (i) an electrostatic potential field in and around a small- or biomacro-molecule. (ii) calculates solvation energy for a single molecule or a group of small molecules, (iii) estimates the binding affinity of a ligand bound to a particular enzyme, (iv) predicts pKa for residues within a protein.

The numerical solution of the electrostatics can be performed by (i) finite difference with considering the neighbouring points (FD), (ii) boundary element (BE) method by using analytical solutions obtained in terms of Green's functions, (iii) finite element (FEM) method is adaptive multilevel approach. It uses tetrahedral elements in the mesh, the dielectric discontinuity is smoothed [20]. The first method is fast with Cartesian mesh and demands low memory, but the resolution of the solution is poor and non-adaptive. The second method smaller numerically and only applicable for linear problems. FEM is highly adaptive and fast [24].

The electrostatic potential map on the van der Waals surface of the Barnase-Barstar protein complex can be seen in *Figure 3.7*.

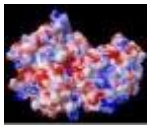


Figure 3.7. Electrostatic potential on the solvent accessible surface of Barnase-Barstar protein (PDB Id.: 2BRS without structural water) calculated by Delphi (red: negative, blue: positive, white: neutral).

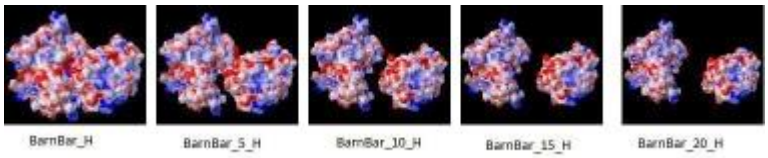


Figure 3.8. Electrostatic potential on the van der Waals surface of Barnase-Barstar protein association at different distances from each other (PDB Id.: 2brs without structural water) calculated by Delphi (red: negative, blue: positive, white: neutral), BarnBar_H: complex, BarnBar_5_H: distance between mass centres is 5 Å, BarnBar_10_H: between mass centres is 10 Å, BarnBar_15_H: between mass centres is 15 Å, BarnBar_20_H: between mass centres is 20 Å.

Figure 3.8. describes the change in electrostatic potential at different distances between the centre of masses.

7.1. Solution of LPBE

The PBE is usually applied to a one to one salt solution and the PBE becomes Eq. 3.26

$$\nabla^2\varphi=\kappa^2\varphi \tag{3.26}$$

(3.29) where $\kappa^2=2z^2eFc_0/(kT\varepsilon_0)$, z is the charge of the ion, F is the Faraday constant, T is the temperature, k is the Boltzmann constant, c_0 is the concentration. Eq. 3.26 is valid for (i) ions with spherical field (Debye-Hückel theory), (ii) ions near a charged plane. The linear PBE (see later) becomes in spherical coordinates (3.27)

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\varphi}{\partial r}\right)=\kappa^2\varphi \tag{3.27}$$

The solution is (3.28)

$$\varphi = A\frac{e^{-\kappa r}}{r} \tag{3.28}$$

where *A* is a constant. *A* can be given for two boundary conditions [25].

8. Langevine and Brownian dynamics

The mathematical basis of molecular motion in a solvent was developed by Paul Langevin. The model of the solvation media is implicit solvation. The basic expression of the Langevin Dynamics (LD) is described in Eq. 3.29 with canonical ensemble. It is a stochastic equation with N particles ($\Sigma_{i=1}i$ = N, the total number of molecules). It does not consider the hydrophobic effects and the electrostatic screening.

$$m_i\frac{\partial^2 \mathbf{r}_i}{\partial t^2} = -\nabla\varphi(\mathbf{r}_i) - \gamma m_i\frac{\partial \mathbf{r}_i}{\partial t} + \sqrt{2\gamma \mathbf{k}_B T m_i}R_i(t) \tag{3.29}$$

where $\varphi(\mathbf{r}_i)$ is the particle interaction potential. The first expression is the particle interaction force. The second and the third expressions in the right side are the frictional and the random force, respectively. The frictional force represents the viscosity of the solution, the random force reepresents the thermal motion of the solvent molecules. *R(t)* is a delta correlated stationary Gaussian process Eq. 3.30-3.31:

$$\langle R(t)\rangle = 0 \tag{3.30}$$

$$\langle R(t)R(t')\rangle = \delta(t-t') \tag{3.31}$$

where δ is the Dirac delta. γ is the friction constant. The greater the γ, the larger the viscosity is. Generally, 5-20 ps^{-1} is chosen in the simulation. The integration time (see Chapter 5) is not a real time. The simulation can give us information on the folding of peptides and small proteins in solvents. The main problem that the real solvent-solute interaction can not be described by this method.

$$0 = -\nabla\varphi(\mathbf{r}_i) - \gamma m_i\frac{\partial \mathbf{r}_i}{\partial t} + \sqrt{2\gamma \mathbf{k}_B T m_i}R_i(t) \tag{3.32}$$

Eq. 3.29 and Eq. 3.32 can be solved by the methods of solution applied in molecular dynamics (MD) (see Chapter 5).

Langevin and Brownian dynamics are good methods for studying folding, association of peptides, proteins and protein-ligands.

Brownian dynamics can be used for the calculation of diffusion constants of proteins and the motion of proteins on nanoscale metal particles with an interface SDA developed by Wade et al. [26].

9. Summary

The electrostatic properties of extended structures, solvation free energies, association energies and the properties of proteins (and other biomolecules) on metal (gold) surface are possible by the solution of Poisson-Boltzmann, Linear Poisson-Boltzmann and Tanford-Kirkwood equations. The folding of peptides, association of proteins can be studied by Langevine or Brownian dynamics.

10. References

1. P. Kukic, J. E. Nielsen, Electrostatics in proteins and protein-ligand complexes, Future Med. Chem. 2(4), 647-666(2010).

2. X. Hao, A. Varshney, "Efficient Solution of Poisson–Boltzmann Equation for Electrostatics of Large Molecules", High Performance Computing Symposium 71–76 (2004).

3. P. A. Atkins, Physical Chemistry, 6th Edition, Oxford, UK, 2004.

4. M. K. Gilson, Introduction to electrostatics, with molecular applications. www.gilsonlab.umbi.umd.edu.

5. K. A. Sharp, B. Honig, Electrostatic interactions in macromolecules: theory and applications. Annu. Rev. Biophys. Chem. 19, 301-332(1990).

6. X. Shi, P. Koehl, The Geometry Behind Numerical Solvers of the Poisson-Boltzmann Equation Comm. in Comp. Phys. 3(5), 1032-1050(2008).

7. B. Z. Lu, Y. C. Zhou, M. J. Holst, J. A. McCammon, Recent Progress in Numerical Methods for the Poisson-Boltzmann Equation in Biophysical Applications. Comm. in Comp. Phys. 3(5) 973-1009(2008).

8. S. S. Kuo, M. D. Altman, J. P. Bardhan, R. Tidor, J. K. White, Fast Methods for Simulation of Biomolecule Electrostatic, Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference, Date of Conference: 10-14 Nov. 2002.

9. M. Holst, F. Saied, Multigrid Solution of the Poisson-Boltzmann Equation, J. Comput. Chem., 14, 105-113(1993).

10. N. Perrin, Probabilistic Interpretation for the Nonlinear Poisson-Boltzmann Equation in Molecular Dynamics, ESAIM: Proceedings, 35, 174-183, March 2012.

11. F. Fogolari, P. Zuccato, G. Esposito, P.Viglino, Biomolecular Electrostatics with the Linearized Poisson-Boltzmann Equation, Biophys. J. 76, 1-16 (1999).

12. J. H. Chaudhry, Stephen D. Bond, Luke N. Olson, Finite Element Approximation to a Finite-Size Modified Poisson-Boltzmann Equation, J. Sci. Comp. 47(3), 347-364(2011).

13. A. Abrashkin, D. Andelman, H. Orland, Dipolar Poisson-Boltzmann equation: ions and dipoles close to charge interfaces, Phys. Rev. Lett. 99(7), 077801(2007).

14. B. Jayaram, D. L. Beveridge, Tanford-Kirkwood Theory for Concentric Dielectric Continua: Application to Dimethylphosphate, Biopolymers, 27, 617-627(1988).

15. F. L. B. Da Silva, B. Jönsson, R. Penfold, A Critical Investigation of the Tanford-Kirkwood Scheme by means of Monte Carlo Simulations, Protein Sci. 10, 14151425(2001).

16. J. J. Havranek, P. B. Harbury, Tanford-Kirkwood electrostatics for protein modeling. Proc. Natl. Acad. Sci. USA 96, 11145-11150(1999).

17. M. Schnieders, J. W. Ponder, Polarizable Atomic Multipole Solutes in a Generalized Kirkwood Continuum, J. Chem. Theory Comput. 3, 2083-2097(2007).

18. C. Tanford, J. G. Kirkwood, Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. J. Am. Chem. Soc., 79 (20), 5333–5339(1957).

19. B.Honig and A.Nicholls. Classical electrostatics in biology and chemistry. Science. 268, 1144-1149 (1995).

20. J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott and J. A. McCammon, Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program, Comp. Phys. Comm. 91, 57-95 (1995).

21. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA 98, 10037-10041(2001).

22. a) D. Bashford, K. Gerwert, Electrostatic Calculations of the pKa Values of Ionizable Groups in Bacteriorhodopsin, J Mol Biol vol. 224, 473-486 (1992).. b) J. L. Chen, L. Noodleman, D. A. Case, D. Bashford, Incorporating Solvation Effects Into Density Functional Electronic Structure Calculations, J. Phys. Chem., 98 (43), 11059–11068(1994). c) D. Bashford, Y. Ishikawa, R. R. Oldehoeft, J. V. W. Reynders, M. Tholburn, An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules Eds, Scientific Computing in Object-Oriented Parallel Environments, volume 1343 of Lecture Notes in Computer Science, pages 233-240, Berlin, 1997.

23. J. A. Grant, B. T. Pickup, A. Nicholls A, A smooth permittivity function for Poisson-Boltzmann solvation methods. J. Comput. Chem., 22, 608-640(2001).

24. J.-P. Hsu, B.-T. Liu, Exact Solution to the Linearized Poisson-Boltzmann Equation for Spheroidal Surfaces, J. Coll. And Interface Sci. 175, 785-788(1996).

25. X. Cheng, Implicit Solvation Models, Introduction to Molecular Biophysics, ÚT/ORNL Center for Molecular Biophysics, 2008.

26. R. R. Gabdouline, R. C. Wade, Brownian Dynamics Simulation of Protein-Protein Diffusional Encounter, METHODS: A Comparision to Methods in Enzymology, 14, 329-341(1998).

11. Further Readings

1. R. Leach, Molecular Modelling, Principles and Applications, 2nd Edition, Prentice Halls, pp. 603-608, Pearson Education Limited, 2001.

2. B.Z. Lu, Y. C. Zhou, M. J. Holst, J. A. McCammon, Recent Progress in Numerical Methods for the Poisson-Boltzmann Equation in Biophysical Application, Review Article, Comm. in Comp. Phys. 8(5), 973-1009(2008).

3. M. J. Holst, The Poisson-Boltzmann Equation. Analysis and Multilevel Numerical Solution. Monograph based on the Ph.D. Thesis below). Applied Mathematics and CRPC, California Institute of Technology, 1994.

4. M. Holst, N. Baker, and F. Wang, Adaptive Multilevel Finite Element Solution of the Poisson-Boltzmann Equation I: Algorithms and Examples. J. Comput. Chem., 21, 1319-1342(2000).

5. J. P. Bardhan, Numerical Solution of Boundary Integral Equations for Molecular Electrostatics, J. Chem. Phys. 130(9), 094102(2009).

6. P. Grochowski, J. Tylska, Continuum molecular electrostatics, salt effects, and countarion binding – a review of the Poisson-Boltzmann theory and its modifications, Biopolymers, 89(2), 93-113(2008).

7. B. Kraczek, Solving the Poisson-Boltzmann Equation, ICES Multi-scale Group meeting, Sept. 29, 2008.

8. W.Rocchia, E.Alexov, and B.Honig. Extending the Applicability of the Nonlinear Poisson-Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. J. Phys. Chem. B 105(28), 6507-6514(2001).

9. B. L. Tembe, J. A. McCammon, Ligand-Receptor Interactions, Comp. Chem. 8(4), 281-283(1984).

10. N. Baker, M. Holst, F. Wang, Adaptive multilevel finite element solution of the Poisson–Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems, J. Comp. Chem. 21, 1343–1352(2000).

11. M. T. Neves-Petersen, S. B. Petersen, Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules-applications in biotechnology, Biotechnol. Annu. Rev. 9, 315-395(2003).

12. Questions

1. What is the interaction between two point charges in a media with ε dielectric constant?

2. Please, write the Coulomb's law.

3. The Poisson equation in a media with constant permittivity.

4. The Poisson equation in a media with variable permittivity.

5. The Boltzmann distribution.

6. The Poisson-Boltzmann equation

13. Glossary

Coulomb's law The electrostatic potential depends ont he point charge and ont he reciproc of the distance.

Poisson equation A function between electrostatic potential and the charge density considering the effective uniform or not uniform dielectric constant.

Boltzmann distribution The distribution of the systems with different energy levels (vibrational, rotational).

Poisson-Boltzmann equation Coupling of Poisson and Boltzmann equations which consideres the effect of ions around the extended structure.

Tanford-Kirkwood equation The solvent excluded spheres are considered in the calculation of electrostatic potentials from the atomic charges.

Van der Waals surface The van der Waals surface of a molecule consideres the van der Waals radius of atoms which build up the molecules and the surface covers the molecules.

Molecular surface with entrant and no-entrant surfaces.

Connolly surface A probe sphere is rolling on the van der Waals surface of the molecule. Th centre of the probe sphere describe a surface – the Connolly surface.

Chapter 4. Solvation Models

(Tamás Körtvélyesi)

Keywords: solvent models, water models, explicit water molecule, implicit solvation models, simple dielectric model, generalized Born model, Poisson-Boltzmann solvation model

What is described here? The biological processes are working in solution. The structure of the solvent can be determined by two main methods: (i) the solvent is considered as individual molecules, or (ii) a space which is similar to the solvent space with the suitable electrostatics.

What is it used for? The solvation models try to simulate the effect of the solvent on the solute to describe the solvent-solute interactions in the reality..

What is needed? The basic knowledge of physical chemistry is necessary on solvents and solutions. The secondary interactions described mathematically with different functions (with and without constrainsed) between the molecules are also important.

1. Introduction

Most of the biological processes take place in solution. The solvent is water at a given pH and ionic strength. In some cases these reactions take plave near membrane. To simulate the structure and the reactions some methods are available: (i) considering the water molecules explicitly [1], (ii) continuum solvation models (simple dielectric models with constant or distance dependent dielectric constants, etc., Poisson-Boltzmann equation (PBE) (see Chapter 3), generalized Born (GB) with SAS (solvent accessible surface) based nonpolar term), continuum dielectric with full treatment of nonpolar solvation [2,3]. In the explicit solvation model a lot of water models were developed. The main problem in many of these models the lack of polarizability of the water molecules which can cause the difference between the model and experiments. There are some methods to correct this difference. The application of the continuum solvation model in molecular mechanics (MM) and molecular dynamics (MD) is faster than in the explicit models.

2. Explicit Solvation Models

The explicit model includes the geometry and charges of the water molecules (see *Figure 4.1.*).

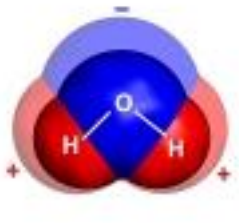


Figure 4.1. One of the explicit water molecule with the partial positive charges on H atoms (δ+) and the partial negative charges on the O-atom (δ-). The total charge is zero

The explicit water models have different geometrics and partial charges on H-atoms and O-atom. The main problem that the fixed charges and the suggested geometry do not reproduce the dipole moment of the water molecule. The reason is the lack of polarizability in most of the water models. SPC and SPC/E models have the same geometry with different partial charges. TIP3P has another geometry and partial charges. TIP4P has an extra dipole moment vector in the mass centre to correct the dipole moment of the model. There are some polarizable water model (e.g. in AMOEBA [4]), but their use are time consuming in MM and MD calculations. We have to decide to make a long simulation with point charge water models (sometimes with ca. 100 thousend water molecules) or a shorter simulation with polarizable water models. In the preparation of MD calculations a flexible water model is used. In the productive simulation rigid water molecules are considered in the periodic boundary condition (PBC) (see Chapter 5). Some physical chemical properties of point charge water models can be seen in Table 4.1.

Table 4.1. Calculated physical parameters of some water models [1]

Model	Dipole moment (μ/D)	Relative dielectric constant (ε_r)	Self diffusion (Dself/(10^{-5} cm^2/s))	Density maximum/ °C	Average config. energy/ (kJ/mol)	Expansion coefficient/ 10^{-4} $°C^{-1}$
SPC	2.27	65	3.85	-13	-45	7.3
SPC/E	2.35	71	2.49	-45	-38	5.14
TIP3P	2.35	82	5.19	-91	-41.1	9.2
TIP4P	2.18	53	3.29	-25	-41.8	4.4
TIP5P	2.29	81.5	2.62	4	-41.3	6.3

Some of the errors in % related to the experimental physical chemical data are summerized in Table 4.2. As it can be seen in some cases significant errors were found in the comparision with the experimental data.

Table 4.2. Errors calculated with rigid water models at 298 K [1] in % of the experimental value

Model	Specific heat capacity (c_p)	Shear viscosity	Thermal conductivity
SPC	102	31	144
SPC/E	108	37	153
TIP3P	107	36	146
TIP4P	118	47	135
TIP5P	120	88	111

The explicit solvent models make possible to simulate biomolecules (peptides, proteins, etc.) in other solvents (dimethyl-sulphoxide, trifluoro-ethanol, dimethyl-formamide, urea, etc.) or in a mixture of organic compounds and water molecules by using explicit solvent models. Before the simulation the mixture have to be equilibrated.

3. Simple models

3.1. Geometric models

The method based on the suppose that the water molecules in the first shell have the main effect (with the solvent accessible surface area of the molecule) on the solvation free energy with its geometry. The main effect of the solvent is its shape and measure. The method is not very accurate, the solvation free energy calculated is not very precise and not depends on the conformational structure. One of the method is the EEF1 [5].

3.2. Dielectric models

In the simple models the water was described as continuum medium [6]. In the simple models the effective dielectric constant can be considered to be constant in the whole system (ε_r=80). On the basis of another approach: near the protein the effective dielectric constant depends on the distance from the protein: distance dependent dielectric constants ε_r= 4r (or ε_r = 4.5 r).

Mehler and Solmajer [7] suggested a sigmoidal dielectric constant dependence on the distance

$$\varepsilon_r(r) = A + \frac{B}{1 + k e^{-\lambda B r}} \tag{4.1}$$

B=ε_r-A, ε_r the effective dielectric constant at 298 K, ε_r = 78.4, A= -88.525, λ= 0.003627, k= 7.7839. This method is used mainly in docking procedure.

Continuum dielectric: solution of Poisson-Boltzmann (PBE), generalized Born (GB) model.

4. Models based on GB/SA and PB/SA

The influence of solvent molecules on the solute is to to transfer the solute from vacuum to water in a given fixed configuration (solvation free energy) considering the free energy of van der Waals interaction, the free energy of cavity formation in the solvent, the free energy of polar to nonpolar structure [8].

The total solvation free energy includes the electrostatic and nonpolar part.

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{nonp} \tag{4.2}$$

Considering the detailed solvation process we can obtain Eq. 4.3

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{vdW} + \Delta G_{cav} \tag{4.3}$$

ΔG_{cav} is the free energy to create a cavity in the solvent.

The nonpolar solvation free energy is based on the solvent accessible surface area

$$\Delta G_{solv} = \sum_i \sum_j \gamma_i S_{ij} \tag{4.4}$$

i and *j* mean the free energy increment (nonpolar solvation free energy in a unit surface) of an atom type (e.g. O, N, etc) and the solvent accessible surface of *j* which is around the atom *i*, respectively. γ is ca. 21 J/(mol Å^2).

4.1. Poisson-Boltzmann method for the calculation of electrostatic solvation free energy

In the Poisson-Boltzmann method, the solution of the non-linear PBE is necessary (see Chapter 3, Eq. 3.21). The solution was performed twice: one for vacuum and one for solution. The difference is the electrostatic free energy of solvation [2,9]:

$$\Delta G_{elec} = \frac{1}{2} \sum_i q_i [\phi_s(\mathbf{r}_i) - \phi_v(\mathbf{r}_i)] \tag{4.5}$$

where ϕ_s and ϕ_v are the electrostatic potential in the solution and in the vacuum, respectively. ϕ is calculated by the finite difference method, which is an expensive calculation and can not be applied in MD calculations, but in molecular mechanics (MM).

The Poisson-Boltzmann equation can be used in molecular dynamics as it was described in Chapter 3.

4.2. Generalized Born method for the calculation of electrostatic solvation free energy

The solution of the non-linear PBE is simplified to use pairwise expressions [10,11]:

$$\Delta G_{elec} = \frac{1}{2}\left(\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_p}\right) \sum_{ij} \frac{q_i q_j}{\sqrt{[r_{ij}^2 + R_i^{GB} R_j^{GB} \exp(-r_{ij}^2 / 4 R_i^{GB} R_j^{GB})]}} \tag{4.6}$$

ε_w and ε_p are the effective dielectric constants of water (ε_w) and the effective dielectric constant in protein (ε_p).

The Eq. 4.6 includes the effective Born radiuses (R_i^{GB} and R_j^{GB}). The values have to satisfy the Born equation:

$$\Delta G_{elec} = \frac{1}{2}\left(\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_p}\right) \sum_{ij} \frac{q_i^2}{R_i^{GB}} \tag{4.7}$$

$R_i^{GB} > R_j^{GB}$, "the effective Born radius is the distance between a particular atom and the effective dielectric boundary" [1]. R^{GB} is a parameter calculated by PBE calculations. The calculations are much more faster than the salvation of PBE, so it is suitable for molecular dynamics (MD) calculations. A lot of version of GB/SA method was developed (see Lit. [2], the Generalized Born Zoo). The main problem is the determination of the solute-solvent boundary: molecular surface (MS) or van der Waals surface (vdW). The previous method is expensive (GBMV), the latter is inexpensive (GBSW). Generally, the effective dielectric constants are 1 and ca. 80 in the protein and in the water, respectively. Methods were developed with variable effective dielectric constants.

5. Summary

The appropriate modeling the molecular properties in solution is basic in the calculations of biomolecules. The explicit and implicit solvation models give a wide range of methods. The explicit salvation method is much more expensive than the implicit solvation method, but the latter method does not cover all of the properties of the solvent (e.g. viscosity).

6. References

1. Martin Chaplin, Water Structure and Science, http://www.lsbu.ac.uk/water/

2. J. Chen, Implicit Solvent, General Principles and Models in CHARMM, Kansas State University, MMTSB/CTBP Workshop, August 4-9, 2009.

3. S. Genheden, P. **Mikulskis**, L. **Hu**, J. **Kongsted**, P. **Söderhjelm**, U. **Ryde**, Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. **J. Am. Chem. Soc.**133(33), 13081-92(2011).

4. a) P. Ren, C. Wu and J. W. Ponder, Polarizable Atomic Multipole-based Potential for Proteins: Model and Parameterization, in preparation. b) P. Ren, C. Wu and J. W. Ponder, Polarizable Atomic Multipole-based Potentials for Organic Molecules, in preparation c) J. W. Ponder and D. A. Case, Force Fields for Protein Simulation, Adv. Prot. Chem., 66, 27-85 (2003). d) P. Ren and J. W. Ponder, Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation, J. Phys. Chem. B, 107, 5933-5947 (2003). e) P. Ren and J. W. Ponder, A Consistent Treatment of Inter- and Intramolecular Polarization in Molecular Mechanics Calculations, J. Comput. Chem., 23, 1497-1506 (2002). f) J. Wang, P. Cieplak and P. A. Kollman, How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calcluating Conformational Energies of Organic and Biological Molecules? J. Comput. Chem., 21,1049-1074 (2000).

5. T. **Lazaridis, M. Karplus , Effective energy function for proteins in solution, *Proteins*,35 (2), 133–52(1999).**

6. D. **Eisenberg, A. D. McLachlan, Solvation energy in protein folding and binding. *Nature*319 (6050): 199–203(1986).**

7. E. L. Mehler, T. Solmajer, Electrostatic effects in proteins: comparision of dielectric and charge models, Proteins Engineering, 4, 903-910(1991).

8. R. M. Levy, L.Y. Zhang, E. Gallicchio, and A.K. Felts, On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute Solvent Interaction Energy. J. Am. Chem Soc., 125, 9523-9530 (2003).

9. F. Fogolari, A. Brigo, H. Molinari, Protocol for MM/PBSA Molecular Dynamics Simulations of Proteins, Biophys. J. 85, 159-166((2003).

10. a) A. Jaramillo, S. J. Wodak, Computational Protein Design is a Challenge for Implicit Solvation Models, Biophys. J. 88, 156-171(2005). b) G. Chopra, C. M. Summa, M. Levitt, Solvent dramatically affects protein structure refinement, PNAS, 105, 20239-20244(2008).

    c) Z. Zhang, S. Wiltham, E. Alexov, On the role of electrostatics on protein-protein interactions, Phys. Biol. 8(3), 035001(2011). d) J. Wang, W. Wang, S. Huo, M. Lee, P. A. Kollman, Solvation Model Based on Weighted Solvent Accessible Surface Area, J. Phys. Chem. 105, 5055-5067(2001).

11. a) J. D. Madura, M. E. Davis, M. K. Gilson, R. C. Wade, A. B. Luty, J. A. McCammon, Biological Applications of Electrostatic Calculations and Brownian Dynamics Simulations, Reviewa in Computational Chemistry V, Ed. by K. B. Lipkowitz, D. B. Boyd, VCH Publishers, New York, 1994. b) W. C. **Still, A. Tempczyk, R. C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, J. Am. Chem. Soc.112 (16), 6127–6129(1990).**

7. Further Readings

1. B. Roux, T. Simonson, Implicit Solvent Models; Biophysical Chemistry; 78; 1999; 1-20.

2. P. Ferrara, A. Caflisch, Folding Simulations of a Three-Stranded Antiparallel Beta-Sheet Peptide; Proc. Natl. Acad. Sci. USA; 2000; 97; 10780.

3. P Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a Fast Implicit Solvent Model for Molecular Dynamics Simulations; Proteins 2002; 46; 24-33.

4. S. J. Wodak, J. Janin, Analytical Approximation to the Solvent Accessible Surface Area of Proteins; Proc. Natl. Acad. Sci. USA; 1980; 77;

5. W. Hasel, T. F. Hendrickson, S. W. Clark, A Rapid Approximation to the Solvent Accessible Surface Area of Atoms; Tetrahedron Computer Methodology 1988; Vol. 1; No. 2; 103-116.

6. F. Fraternali, W. F. van Gunsteren, An Efficient Mean Solvation Force Model for Use in Molecular Dynamics Simulations of Proteins in Aqueous Solution; J. Mol. Biol. 1996; 256; 939.

8. Questions

1. What kind of main models can describe the solution models?

2. What is the difference between the explicit solvent models?

3. What are the main problems with the explicit solvent models?

4. Please, give some examples on the simple models!

5. Please, describe the generalized Born solvation method!

6. Please, describe the PBSA solvation method with MM!

9. Glossary

Explicit solvation models: Solvent model with the individual solvent molecules as models.

Geometric model: It considers the geometric models of the solvents in the interactions of solvent molecules and solvent-solute molecules.

Generalized Born model: A pairwise description of the electrostatic interactions in water by using the simplified PBE.

MM/Poisson-Boltzmann model: Solution of the PBE with MM.

Chapter 5. pK_A Calculations of Biologically Active Molecules

(Tamás Körtvélyesi)

Keywords : pK_A shift of aminoacid side-chains in peptides and proteins, prediction the protonation (pK_A) of side chains in peptides and proteins.

What is described here? The biologically active molecules are working in solution which has *pH* and ionic strength (*I*). The *pH* and ionic strength (*I*) have influence on the charges in the side chains of the amino acids of peptides and proteins. The charges and stability of DNA are also sensitive on the *pH* and *I*. There are two main methods to predict the pK_A : (i) empirical calculation of the pK_A shift in peptides and proteins, (ii) non linear Poisson-Boltzmann equation (NPBE), linear Poisson-Boltzmann equation (LPBE) calculations with finite different method and the solution of the Tanford-Kirkwood equations (TKEs) combined with Monte Carlo methods to obtain the shift of the pK_A in the molecules.

What is it used for? The prediction of pK_A is important to know the protonation state of the side chains in peptides and proteins to predict the electrostatic interactions between the peptides, proteins and ligand molecules in the possible association processes.

What is needed? The basic knowledge of the structure, intra- and intermolecular interactions between molecules are important. Important also, the knowledge of the introduction to physical chemistry. The basic analytical chemistry knowledge on the acidity is also necessary.

1. Introduction

The pK_A of the side chains in peptides or proteins of aminoacids at a given *pH* are significant in the folding, in the formation of the structures and the working of protein-protein and protein-ligand binding. These protonation states affect the structure and stability of the peptides and proteins and also the binding mode of the pocket in these molecules. The pK_A values in peptides and proteins can be derived from the protonation constants of aminoacids alone.

The peptides and proteins fold in solvent and stabilize the 3D structure on the basis of the effect in the side chain charges (protonation). The titratable/hydrophylic amino acids are on the surface of the water/protein interface. In the core of the peptide/protein, hydrophobic aminoacids are burried. The pK_A values of the side chains in aminoacids are summerized in Table 5.1. (Henderson-Hasselbach equation must be checked in Lit. [1], Eq. 6.1)

Table 5.1. The pK_A values of side chains in individual aminoacids

Asp	3,9
Glu	4,1
His	6,0
Cis	8,4
Tyr	10,5
Lys	10,5
Arg	12,5

The charged side chains are in interction with each other and with the backbone atoms by point charge-point charge, point charge-dipole and dipole-dipole, etc. interactions. The effect on the pK_A values of the side chains depends on (i) pH and (ii) independent on pH. The latter influences are the desolvatation, interactions with the constant charges and dipoles. It means that the pK_A values change in peptides/proteins related to the individual aminoacids. The *pH* dependent part can be determined by Tanford-Roxby iteration or other methods (see later). With the determination of the pK_A values the titration curves of peptides/proteins can be calculated (Henderson-Hasselbalch titration curves if all the charged side chains behave on the basis of Henderson-Hasselbalch, etc.)

2. Empirical Methods

PropKa

One of the empirical method which has no potential calculations is Propka [2-5]. The method is very fast. Version 1.0 [2] predicts pK_A of protein on the basis of the charge groups distances and empirical formulas. In Version 2.0 [3] a new function was developed to predict the effect of non protein molecules (ligands) on pK_A of the side chain protonation of protein and the shift of *pK* of ionizable groups on ligand. Ligands and the aminoacids with charged side chain far from the binding site can contribute protonation/de protonation. The small change in sequence can shift the pK_A. The parameters and the empirical rules were changed in the new version [4]. This method describe more precisely the desolvatation and dielectric response of the proteins. The classification of aminoacids was modified to internal aminoacids and amino acids on the surface. The method is precise to pK_A of Asp and Glu. The newest method (Version 3.1 [5]) includes the effect of the ligand ont he protonation state of the binding site and the *pK* shift of the ionizable nonprotein ligands. It makes possible th calculate the shift in multiligand complexes and non covalent coupled ligands which were not possible previously. The database was renewed and can be extended flexibly. A GUI was developed for the simple usage of the method [6]. The method is implemented in a web server [7].

In *Figure 5.1*. the XRD apo structure and the structure without its ligand (geldanomycine) can be seen to describe the differences which is dues to the induced fitting of the ligand. At pH=7.2 only the default protonation is valid. If the concentration of the protonated side chains are more than 50% than we accept as protonated side chain (see *Figure 5.2*).

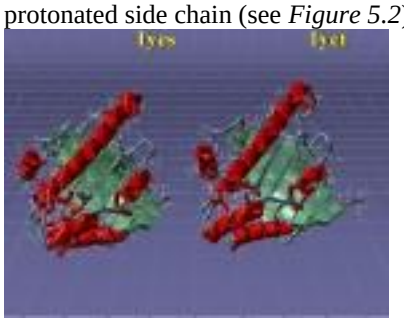


Figure 5_1. HSP90 N-terminal structures: apo structure (1yes) and structure without the ligand (geldanomycine) (1yet)

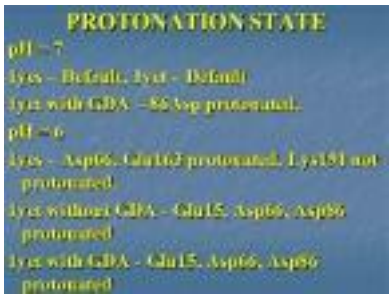


Figure 5.2. The protonation of HSP90 N-terminal structures: apo structure (1yes) and structure without the ligand calculated by propKa 3.(geldanomycine) (1yet)

3. Solvation of Poisson-Boltzmann Equation (PBE) and the Tanford-Kirkwod Equations (TKE) Coupled with Monte Carlo Methods

The application of the solution of Poisson-Boltzmann Equation (PBE)

The shift in pK_A in peptides/proteins depend on two factors: (i) the electrostatic environment of the charged (protonated/deprotonated) side chains of the amino acid, and the embedding of the side chain. These factors are influenced by the geometry of the molecules and *vica versa*.

Some methods are based on the solution of PBE mainly by FDPB (finite difference Poisson-Boltzmann method) or LPDF (linear Poisson-Boltzmann method) (see Chapter 3). It includes the modifications of the electrostatic environment int he protein. Some web servers are installed: H++ web server [8], pK_D web server [9], az MCCE [10] and the Karlsberg+ FDPB [11] method. The Tanford-Kirkwood equation is solved in Macrodox [12]. The latter method is suitable for Brownian dynamics, too.

The FDPB-based methods support the pK_A shift of the amino acid side chains with the difference of the totally solvated and in-protein condition. It is necessarry to give the effective dielectric constant in protein and int he bulk solvent. The previous value is between 2 to 20 (see Chapter 3).

H++ method

The H++ server predicts the pK_A values of the ionizable side chains and automatically extend/delete the protons [8,13]. The input file has to be in pdb format, the output will be in pdbq, pqr format, pdb format or amber format. The theoretical background is summerized in Lit. [14]. We can calculate the isoelectronic points, titration curves and the protonation microstates. The titration curve of the whole protein and separately, the charged groups can be obtained.

Karlsberg+ method

On the basis of the LPBE solution and the structural relaxation of H-atoms and salt bridges the pK_A values are calculated [11,15]. The effective dielectric constant in the protein is 4.0. The method is capable to predict at different pHs the conformation and the position of H-atoms. The pH dependent conformations at different protonation states are calculated by Monte Carlo simulations. LPBE calculations were performed by TAPBS algorithm.

Karlsberg+ can calculate with optimization the electrostatic energies of the conformers of the proteins. The H-atoms ont he surface of the protein and the salt bridges are calculated at three pHs (low, middle and high) [15]. The pdb file of the protein is necessarry.

Macrodox method

Macrodox [12] is a command line pakage on Linux and Windows XP environment The algorithm uses the solution of Tanford-Kirkwood-eqation. In this calculation we can obtain the protonation/deprotonation states which depends on pH and ionic strength, change in temperature. The effective dielectric constant of the solvent, The internal dielectric constantat ionic strength of the media, pH can be changed. The command „titrate" starts the calculations. Protein is considered as a substance with low effective dielectric constant (we do not know precisley on the real effective constant in the internal part of the protein which is in the solvent with high effective dielectric constant. The effect of the buried ionazable side chains (burial factor) and the interactions between these side chains is important. A good example is the binding pocket in BACE (1fkn), where the Asp int he pocket is protonated which has an important effect on binding molecules (see the binding pocket in *Figure 5.3*). Another effect is the position of loop which depends on pH. The lower the pH is the loop is more open. The local pH in the cells is not the physiological pH, sometimes more or less of this value.

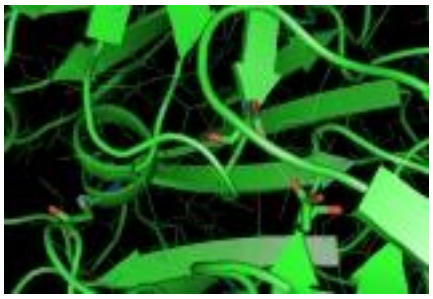


Figure 5.3. Binding pocket in .BACE (1fkn)

The difference maximum in the Barnbar-Barnase protein complexes at 0, 5, 10, 15 and 20 Å distances between the mass centres (see Chapter 3) calculated by different methods are summarized in *Table 5.2*. Mainly, Tyr and His have the largest difference in the calculations. His has two tautomers and one positively charged protonated form. The environment parameters were 298.15 K, effective dielectric constant is 78.3, the internal effective dielectric constant is 4.0, the ionic strength was 0.1 M, the effective radius of the protein is 20.50 Å.

Table 5.2. Largest difference maximum in the Barnbar-Barnase protein complexes at 0, 5, 10, 15 and 20 Å distances between the mass centres (see Chapter 3) calculated by different methods without ligands.

	VegaZZ*	H++	Macrodox	Propka 3.0.	Karlsberg +
	Average of difference maximum	Average of difference maximum	Average of difference maximum	Average of difference maximum	Average of difference maximum
BarnBar	His (7,410)	His (6,554)	Glu (1,514)	Tyr (1,964)	Asp (3,727)
BarnBar_5	His (6,373)	Tyr (3,397)	Glu (1,567)	Tyr (1,197)	His (3,133)
BarnBar_10	His (5,490)	Tyr (3,890)	Glu (1,584)	Tyr (0,936)	Tyr (3,356)
BarnBar_15	His (5,100)	Tyr (3,873)	Glu (1,574)	Tyr (0,926)	Tyr (2,775)
BarnBar_20	His (5,040)	Tyr (3,894)	Glu (1,585)	Tyr (0,920)	Tyr (2,750)

*PropKa 2.0

Table 5.2 supports that the results of the differences are significant at His. His has two tautomers and a protonated form. The differenc in ionizable side chains are much more smaller.

Molecular dynamics calculations

In molecular dynamics calculations the free energy difference between the protonated and deprotonated form are performed. It is possible by free energy perturbation, thermodynamics integration, Bennett transfer ratio and LIE (Linear Interaction Energy) (see Chapter 9). These methods are expencive computationally than the PBE methods. They consider only point charges (see Chapter 2) without polarizability. Most of the pK_A calculation methods suppose the validity of Henderson-Hasselbalch titration curve which means that we can conclude on the pK_A by the half protonation state. Some methods are described int he next section.

The calculations are possible in molecular dynamics in constant *pH*. At given integration steps the pK_A values are calculated and the protonation of the side chains are corrected [16, 17]. In a lot of cases proteins can be obtained commercially in buffers. In their experimental titration must be carefully performed, the results have to accept with critics!

4. Summary

The knowledge of the pK_A values (in aminocid side chains) is important in modelling peptides and proteins. The protonation of the side chains in amino acids have effect on the structure of the moleule and its interactions with ligands to bin din the binding pocket.

5. Acknowledgement

The author is grateful for Krisztina Laskay, who summarized the methods in her B.Sc. Thesis.

6. References

1. P.W. Atkins, Physical Chemistry, Fourth Ed., Oxford University Press, 1990.

2. H. Li, A. D. Robertson, J. H. Jensen, **Very Fast Empirical Prediction and Interpretation of Protein pKa Values**, Proteins, 61, 4, 704-721 (2005).

3. D. C. Bas, D. M. Rogers, J. H. Jensen, **Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes** Proteins, 73, 3, 765-783 (2008).

4. M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, J. H. Jensen, **PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa predictions**, J. Chem. Theory Comput., 7, 2, 525-537 (2011).

5. C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, and J. H. Jensen, **Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values**, J. Chem. Theory Comput., 7, 7, 2284-2295 (2011).

6. M. Rostkowski, M. H.M. Olsson, C. R. Søndergaard and J. H. Jensen **Graphical Analysis of pH-dependent Properties of Proteins predicted using PROPKA**, BMC Structural Biology 2011 11:6.

7. http://propka.ki.ku.dk/

8. **http://biophysics.cs.vt.edu/**

9. **B. M. Tynan-Connolly** and **J. E. Nielsen**, pKD: re-designing protein p***K***a values, **Nucleic Acids Res. 34(Web Server issue): W48–W51 (2006).**

a) G. R.E., Alexov E.G., Gunner M.R., Combining conformational flexibility and continuum electrostatics for calculating pKa's in proteins. Biophys J. 83, 1731-1748(2002). b) E. Alexov and M.R. Gunner, Incorporating protein conformational flexibility into pH- titration calculations: Results on T4 Lysozyme. Biophys. J. 74, 2075-2093 (1997).

10. **http://agknapp.chemie.fu-berlin.de/karlsberg/**

11. http://iweb.tntech.edu/macrodox/mdxhelp/overview.html.

12. a) R. Anandakrishnan, B. Aguila, A. V. Onufriev, H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation, Nucleic Acids Res., 40(W1):W537-541. (2012). b) J. Myers,G. Grothaus, .S. Narayanan, A. Onufriev, A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecule, Proteins, 63, 928-938 (2006). c) J. C. Gordon, J. B. Myers, T. Folta, V. Shoja,L. S. Heath, A. Onufriev. H++: a server for estimating pKas and adding missing hydrogens to macromolecule", Nucleic Acids Res. Jul 1;33:W368-71(2005).

13. D. Bashford, pKa of Ionizable Groups in Proteins, Atomic Detail from a Continuum Electrostatic Model. by D. Bashford and M. Karplus; Biochemistry, 29 10219-10225(1990).

14. G. Kieseritzky, E.-W. Knapp, Optimizing pKA computation in proteins with pH adapted conformations (PACs), Proteins, 71, 3, 1335-1348 (2008). b) B. Rabenstein, E.-W. Knapp, Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers. Biophys J, 80, 3, 1141-1150 (2001).

15. S. Donnini, F. Tegeler, G. Groenhof, H. Grubmüller, Constant pH Molecular Dynamics in Explicit Solvent with lambda-Dynamics. *J. Chem. Theory and Comp.* 7, 1962-1978 (2011).

16. J.Morgan, D. A. **Case**, J. A. **McCammon**, Constant pH Molecular Dynamics in Generalized Born Implicit Solvent, J. Comp. Chem. 25, 2038-2048(2004).

7. Further Readings

1. http://en.wikipedia.org/wiki/Protein_pKa_calculations.

8. Questions

1. Please, describe the main methods to obtain the titration curves of a protein!
2. What influences are considered in the empirical method of pK_A method?
3. Please, give the general expression to the Henderson-Hasselbalch titration curve!
4. Please give the equilibrium equation for acids and basis!
5. Please, write the Poisson-Boltzmann equation!
6. Is it possible to make calculations in constant *pH*?

9. Glossary

Empirical pK_A calculations: The pK_A values (pK_A value shifts from the standard pK_As) are calculated on the basis of the side chains in the neighbourhood and with considering the distance between the ionizable side chains.

pK_A calculations: pK_A calculations can be performed by empirical and PBE/TKE solution combined by Monte-Carlo method.

Titration curves: The pH curve we obtain by the titration of the proteins with charged side chains.

Chapter 6. Molecular Dynamics

(Ferenc Bogár)

Keywords: molecular dynamics, Newton's equation of motion, numerical integration, statistical physics, statistical ensembles, NPT, NVT, thermostat, barostat, constraints, simulated annealing, replica exchange molecular dynamics

What is described here? In molecular dynamics simulations Newton's equation of motion is solved numerically for the atoms of a molecular system (*e.g.* protein in water). With these simulations, we obtain typically a statistical equilibrium ensemble and from this we can calculate, among others, thermodynamical quantities (*e.g.* pressure, energy) or structural informations (like the average helical content of a peptide, which is useful, for example in the interpretation of CD spectra).

What is it used for? MD is one of the most popular methods in biomolecular modelling. It can be used for conformational analysis, structural stability investigations as well as structural transition studies (*e.g.* protein folding studies). It is often exploited in combination with other methods, *e.g.* binding free energy calculations.

What is needed?

- The fundamentals of classical mechanics
- Classical description of molecular forces (Molecular mechanics, Chapter 2)
- Basic knowledge on numerical solution of differential equations
- Basics of statistical thermodynamics
- Basics of calculus

1. Introduction

In molecular systems, at any level, from water molecules to biological macromolecules (like DNS or proteins) the chemical bond plays the central role. This is a non-classical phenomenon and undoubtedly the quantum mechanics is the proper level of theory which is necessary for its description. With the solution of the Schrödinger equation we can account for the formation or breaking of chemical bonds. The solution of this equation, even with approximations, is possible only for small systems. If we want to treat larger systems computationally, we need further approximations. One possibility is to use classical description of the interactions (*i.e.* molecular mechanics) instead of quantum mechanics. The price, we have to pay, is high: this theory is unable to account for the changes in chemical structure. On the other hand, a lot is gained: we can use the Newton's equations instead of Schrödinger's equation.

This simplification enables us to model the molecular system at non-zero absolute temperature using the machinery of the statistical mechanics. The state of the system in this classical model is determined by the positions and momenta of the atoms. At every finite temperatures these states have a characteristic probability distribution, knowing this we can calculate several physical and chemical properties of the system. The determination of the complete distribution would be an enormous task and it is impossible for biological systems. Instead, we use the methods of statistical physics to sample those states that are reachable by our molecules under predefined physical conditions. This sampling can be done using molecular dynamics (MD). In this chapter we describe the basics of this broad and fast developing field of molecular modelling.

2. Fundamentals of molecular dynamics

2.1. Selection of the model system: Cluster calculation or periodic boundary conditions

Although, extremely large simulations (say 10^5atoms) can be carried out on the computers available today, the treatable system size is considerably smaller than the typical amount of material participate in

chemical/biochemical processes (say ~10^{23} atoms). If we simply take our simulated system, it will have a boundary (*e.g.* water-vacuum interface) where the system is in vacuum. The relative size of the surface of this interface is larger than in realistic case, therefore this kind of simulation may overemphasize the surface effects. This can be misleading, except if we want to investigate small *cluster*s (*Figure 6.1. A*) where this phenomena plays a central role. The optimal solution would be to simulate considerably larger systems but it not possible today.

We can also borrow the method of '*periodic boundary conditions*' (PBC) from solid state physics. We select a 3D geometrical figure such that with non-overlapping repeats of if we can completely cover the space (in the simplest case it is a cube). We fill up this object with our system (say a biomolecule and water), this will be the *reference cell*. We shift this reference cell in three directions and cover the whole space with their copies. During the simulation it is required that an atom in the reference cell and its images do the same motion (*Figure 6.1. B* , this figure based on snapshots taken from Democritus MD tutorial program). This is a constraint and the system built up this way is not completely equivalent to an infinite free system but this method eliminates the unwanted surface effects and makes the simulations more realistic. The motion of the particles in the reference cell in a PBC simulation is presented in *Figure 6.1C*.

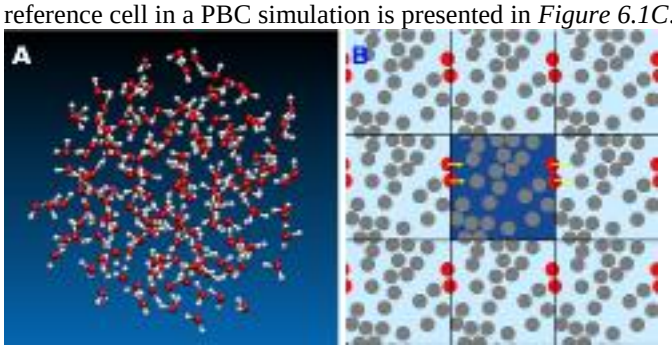


Figure 6.1. A: Cluster of water molecules in vacuum, B: Periodic boundary condition in 2D: The repeat unit (square) is shown in dark blue. (Based on snapshots taken from Democritus: http://www.compsoc.man.ac.uk/~lucky/Democritus/Experiments/exps.html). If one particle leaves the box at one side an other enters at the other side (see the red spots with arrows).

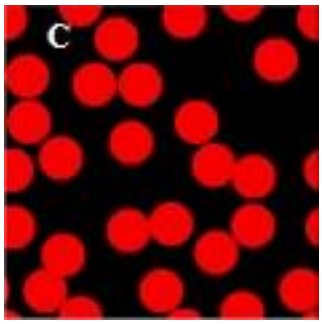


Figure 6.1C. Motion of the particles in a molecular dynamic simulation using periodic boundary conditions. This movie was made with the Democritus program.

2.2. Newton's equation of motion for molecular systems

In molecular mechanics the molecules are considered as mass points with bonded and non-bonded interactions between them (see Chapter 2). The time evolution of a system with N atoms is described by Newton's equation of motion:

$$\mathbf{F}_i = m_i \mathbf{a}_i, \quad \text{with} \quad i = 1, \ldots, N; \tag{6.1}$$

where $\mathbf{a}_i$ is the acceleration, m_i is the mass of the *i*-th atom and $\mathbf{F}_i$ is the force acting on it. $\mathbf{a}_i$ is given by

$$\mathbf{a}_i = dv_i/dt = d^2\mathbf{r}_i/dt^2$$

that is the acceleration is the first derivative of the velocity ($\mathbf{v}_i$) and second derivative of the position ($\mathbf{r}_i$) of the i-th atom. Eq. 6.1 is a system of second order ordinary differential equations with N members. From the theory of

ordinary differential equations we know that we need initial conditions for their solution, namely N positions at the starting time ($\mathbf{r}_{i0}$) and the same number of initial velocities ($\mathbf{v}_{i0}$).

2.3. Calculation of forces

In order to set up the Newton's equation of motion the force acting on the i-th particle need to be defined. This is done using classical forces as it is described by molecular mechanics (see Chapter 2). The largest and most time consuming part of the force calculation in an extended 3D system is the computing of non-bonded pair interactions, as it scales as the second power of the number of atoms in the system. To reduce the computational time often a cut-off radius R_{cut} is introduced and the pair interactions are neglected, if the atoms are farther than R_{cut} from each other.

If we use periodic boundary conditions and the cut-off distance is chosen too large, artefactual interactions may appear (*e.g.* a biomolecule interacts with its counterpart in the neighbouring cells). In order to exclude this effect we use the so called *minimum image convention*: during the calculation of the pair interaction of atom A with atom B, we take always that image of B which is the closest to A. In practice it means that the cut-off radius is chosen as at most the half of the smallest diameter of the repeat unit (*Figure 6.2.*).

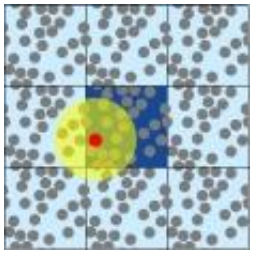


Figure 6.2. Schematic representation of the minimum image convention

2.4. Integration methods

Newton's equation of a large molecular system is solvable only numerically. In numerical integration methods we start from the initial state (position and velocity of atoms) and using a proper time step we generate the solution of the Eq. 6.1, stepwise. We present here three widespread methods (Verlet, leapfrog and velocity Verlet) which are often used in popular MD programs.

Verlet integrator method [1]

Let us suppose that we already know the position vectors and their necessary derivatives at the time of t_k. Using the Taylor expansion of the position vector we can calculate its value at $t_k+\Delta t$ and $t_k-\Delta t$

$$\begin{aligned} \mathbf{r}_i(t_k+\Delta t) &= r_i(t_k) + \left.\frac{d\mathbf{r}_i(t)}{dt}\right|_{t=t_k}\Delta t + \frac{1}{2}\left.\frac{d^2\mathbf{r}_i(t)}{dt^2}\right|_{t=t_k}\Delta t^2 + \frac{1}{6}\left.\frac{d^3\mathbf{r}_i(t)}{dt^3}\right|_{t=t_k}\Delta t^3 + O \\ \mathbf{r}_i(t_k-\Delta t) &= r_i(t_k) - \left.\frac{d\mathbf{r}_i(t)}{dt}\right|_{t=t_k}\Delta t + \frac{1}{2}\left.\frac{d^2\mathbf{r}_i(t)}{dt^2}\right|_{t=t_k}\Delta t^2 - \frac{1}{6}\left.\frac{d^3\mathbf{r}_i(t)}{dt^3}\right|_{t=t_k}\Delta t^3 + O \end{aligned} \tag{6.2}$$

where $O(\Delta t^4)$ is an error term of order t^4. Adding these two series, the terms with odd orders will cancel and we obtain

$$\mathbf{r}_i(t_k+\Delta t)+\mathbf{r}_i(t_k-\Delta t) = 2\mathbf{r}_i(t_k) + \left.\frac{d^2\mathbf{r}_i(t)}{dt^2}\right|_{t=t_k}\Delta t^2 + O(\Delta t^4) \quad . \tag{6.3}$$

After rearrangement

$$\mathbf{r}_i(t_k+\Delta t)=2\mathbf{r}_i(t_k)-\mathbf{r}_i(t_k-\Delta t)+\left.\frac{d^2\mathbf{r}_i(t)}{dt^2}\right|_{t=t_k}\Delta t^2+O(\Delta t^4)\ . \tag{6.4}$$

Using the Newton's equation we can substitute the acceleration with the force calculated from the position vectors at t_k:

$$\mathbf{r}_i(t_k+\Delta t)=2\mathbf{r}_i(t_k)-\mathbf{r}_i(t_k-\Delta t)+\frac{\mathbf{F}_i(t_k)}{m_i}\Delta t^2+O(\Delta t^4) \tag{6.5}$$

In this method we need to know the values of the position vectors at t_k and t_k - Δt (green points in *Figure 6.3.*). The forces at t_k can be calculated from $\mathbf{r}_i(t_k)$-s (blue spot in *Figure 6.3.*). Finally, from the positions and the forces we can calculate the new positions (red spot in *Figure 6.3.*). Moving the reference time one step further we can repeat the procedure until reaching the desired simulation time.

The *Verlet method* does not require the explicit calculation of the velocities but we may need it during the evaluation of the results (*e.g.* calculation of kinetic energy). Formally we can obtain it by subtracting the second equation of (6.2) from the first:

$$\left.\frac{d\mathbf{r}_i(t)}{dt}\right|_{t=t_k}=\frac{1}{2\Delta t}[\mathbf{r}_i(t_k+\Delta t)-\mathbf{r}_i(t_k-\Delta t)]+O(\Delta t^2) \tag{6.6}$$

The error of this approximate value is second order in Δt, considerably larger that in the case of position (fourth order in Δt) .

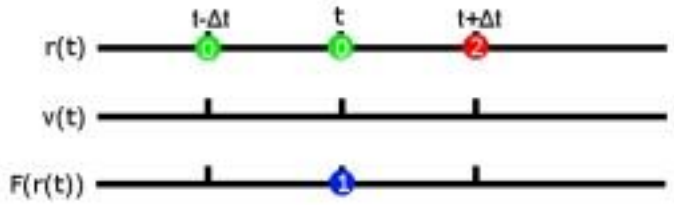


Figure 6.3. Steps of the Verlet integration algorithm

Leapfrog integrator [2]

The other widespread method for the numerical solution of the Newton's equations is the leapfrog integration. It can be easily derived by rearranging the central formula of Verlet's procedure **[6.4]** and dividing it by Δt:

$$\frac{\mathbf{r}_i(t_k+\Delta t)+\mathbf{r}_i(t_k)}{\Delta t}=\frac{\mathbf{r}_i(t_k)-\mathbf{r}_i(t_k-\Delta t)}{\Delta t^2}+\frac{\mathbf{F}_i(t_k)}{m_i}\Delta t^2+O(\Delta t^4) \tag{6.7}$$

The first term in the left hand side and right hand side of the equation is an approximation of $\mathbf{v}_i(t+\frac{1}{2}\Delta t)$ and $\mathbf{v}_i(t-\frac{1}{2}\Delta t)$, respectively. This gives the first equation of the leapfrog method.

$$\begin{aligned}&\mathbf{v}_i(t+\tfrac{1}{2}\Delta t)=\mathbf{v}_i(t-\tfrac{1}{2}\Delta t)+\frac{\Delta t}{m_i}\mathbf{F}_i(t)\ ,\\&\mathbf{r}_i(t+\Delta t)=\mathbf{r}_i(t)+\Delta t\,\mathbf{v}_i(t+\tfrac{1}{2}\Delta t)\ ,\qquad i=1,\ldots,N\end{aligned} \tag{6.8}$$

The second equation comes from the approximate expression of the velocity used in the derivation. The schematic representation of this method is given in *Figure 6.4.*

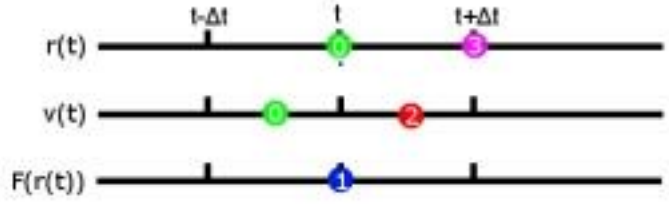


Figure 6.4. Steps of the leapfrog integration algorithm

Velocity-Verlet [3]

The position vector after the time step can be directly calculated from the Taylor expansion (first equation of (6.2)). Substituting the acceleration with the forces, using the Newton's equation

$$\mathbf{r}_i(t_k+\Delta t)=\mathbf{r}_i(t_k)+\left.\frac{d\mathbf{r}_i(t)}{dt}\right|_{t=t_k}\Delta t+\frac{1}{2}\frac{\mathbf{F}_i(t_k)}{m_i}\Delta t^2 \tag{6.9}$$

To obtain this we need the position, velocity and force values at t_k. But for the next step we also need the velocity at t_k that time, $\mathbf{v}_i(t_k + \Delta t)$. This can be calculated as

$$\mathbf{v}_i(t_k+\Delta t)=v_i(t_k)+\int_{t_k}^{t_k+\Delta t}\mathbf{a}_i(t)\,dt\approx v_i(t_k)+\frac{\Delta t}{2}(a_i(t_k)+a_i(t_k+\Delta t))\ . \tag{6.10}$$

Here we used linear approximation of the acceleration in the time interval of (t_k, $t_k + \Delta t$). Using again the Newton's equation we obtain:

$$\mathbf{v}_i(t_k+\Delta t)=v_i(t_k)+\frac{\Delta t}{2m_i}(\mathbf{F}_i(t_k)+\mathbf{F}_i(t_k+\Delta t)) \tag{6.11}$$

The steps of the solution of these equations can be seen in *Figure 6.5*.

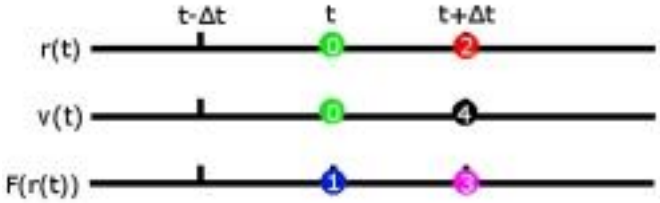


Figure 6.5. Steps of the velocity-Verlet integration algorithm

3. Statistical mechanics background

3.1. Microstates, macrostates

An extended molecular system (*e.g.* protein in a solvent) has a very large number of degrees of freedom. Its states in classical mechanics can be described by the coordinates (**r**) and momenta (**p**) of the *N* particles the system consists of. Every possible combination of these vectors describes a *microstate* ($\mathbf{r}_1$, $\mathbf{r}_2$, . . ., $\mathbf{r}_N$, $\mathbf{p}_1$, $\mathbf{p}_2$, . . ., $\mathbf{p}_N$). The possible microstates form together the *phase space*. A probability distribution $\mathbf{P}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N)$ of the microstates defines a *statistical ensemble*. In biomolecular simulations we only use ensembles related to an equilibrium state of the system under given macroscopic environmental constraints (*e.g.* fixed volume or pressure).

In statistical physics our goal is to derive the macroscopic properties (*macrostates*) of a complex physical system, knowing the particles as well as their interactions in it . To reach our goal we have to know theoretically all of the microstates of the system together with their probabilities (*i.e* .P). In general it is inevitable to know all these data, in most cases we have to settle for the collection of a representative sample that gives a good approximation of the total ensemble.

The central problem of the simulation is, how we can produce a proper approximation of an ensemble using a limited sampling time. Before discussing this problem, we have to mention the problem of ergodicity. Our system is *ergodic* if a single copy of the system will go through all of its microstates, if we follow its evolution (*trajectory*) in the state space for an appropriately long time. Unfortunately this time can also be infinite. In practical simulations we have to find a proper sampling of the ensemble that provides approximate values for the macrostates (averages) which are close enough to the exact values. One of key questions of the simulation is:

How can we test the accuracy of this approximation if we do not know the exact values? What can be done is to calculate the value of a selected property and test its convergence as the simulation time grows (of course, this can also be problematic, because nothing guarantees that the convergence is uniform).

3.2. Ensembles: NPT, NVT, micro canonical, canonical

The probabilities of the microstates depend on the macrostate of the system, which is determined (according to the classical thermodynamics) by the thermodynamic variables. The most often used *conjugate variable pairs* are the entropy/temperature (S/T), volume/pressure (V/p), particle number/chemical potential (N,μ). The first members of these pairs are *extensive*, while the second ones are *intensive* quantities. Fixing any member of the three conjugate pairs we obtain a specific ensemble. The most often used ones are the NVE (microcanonical), NVT (canonical) and NPT (isotherm-isobar) ensembles. The thermodynamic state of these ensembles are defined by fixing quantities listed in the name of the ensemble (*e.g.* N: particle number, V: volume and E: energy).

3.3. Probability distribution in microcanonical, canonical ensembles

In a *microcanonical ensemble* the system is isolated from its environment, neither material nor energy transport is allowed. In this case each microstate has equal probability.

If our system, with fixed particle number and volume, is in equilibrium with a heath bath (which allows the energy exchange between the system and its environment) it is termed as *canonical ensemble*. Its states follow the Maxwell-Boltzmann statistics that is the probability of finding the system in a dV=(d**r**,d**p**) volume element at (**r**,**p**)= (**r**$_1$, **r**$_2$, . . ., **r**$_N$, **p**$_1$, **p**$_2$, . . ., **p**$_N$) in the state space is

$$\mathrm{P}(r,p)\,dr\,d\,\mathrm{p}=\frac{1}{Z}\exp\left(-\left[\sum_i^N\frac{\mathrm{p_i^2}}{2m_i}+U(r)\right]\right)dr\,d\,\mathrm{p}\quad , \tag{6.12}$$

where U(r) in the potential energy of the system,

$$Z=\iint\exp\left(-\left[\sum_i^N\frac{\mathrm{p_i^2}}{2m_i}+U(r)\right]\right)dr\,d\,\mathrm{p} \tag{6.13}$$

is the *partition function*.

3.4. Calculation of ensemble averages

In statistical physics the average of physical quantities f(r,p) can be calculated using the above defined probabilities:

$$\langle\mathrm{f}(r,p)\rangle=\iint\mathrm{f}(r,p)\mathrm{P}(r,p)\,dr\,d\,\mathrm{p}\quad . \tag{6.14}$$

We mention here an alternative formulation of the ergodicity: the ensemble average of an arbitrary quantity is equal to its time average, *i.e.*

$$\langle\mathrm{f}(r,p)\rangle=\frac{1}{Z}\iint\mathrm{f}(r,p)\mathrm{P}(r,p)\,dr\,d\,\mathrm{p}=\lim_{t\to\infty}\frac{1}{t}\int_0^t\mathrm{f}(r(\tau),p(\tau))\,d\tau\quad , \tag{6.15}$$

where **r**(τ) and **p**(τ) denote the positions and momenta of the atoms at the time τ, respectively.

Averaging using trajectories from equilibrium MD simulations

Having proper sample of the statistical ensemble of our system the averages can be calculated directly. From the position and momentum vectors we can calculate the value of f_i (r,p) the physical quantity *f* at the i-th time point of the trajectory. If we have altogether M points along the trajectory the average is

$$\langle \mathrm{f} \rangle = \frac{1}{M} \sum_{i=0}^{M} \mathrm{f}_i(r,p) \quad . \tag{6.16}$$

The standard deviation of the quantity is

$$\Delta \mathrm{f} = \langle (\mathrm{f} - \langle \mathrm{f} \rangle)^2 \rangle^{1/2} \quad . \tag{6.17}$$

3.5. Examples:

Calculation of temperature

As we have learned from statistical physics the temperature of molecular system containing N particles can be calculated from the average kinetic energy and the equipartition theorem .

$$\langle \mathrm{K} \rangle = \frac{3}{2} N k_B T \, . \tag{6.18}$$

Here k_B is the Boltzmann constant, T is the absolute temperature.

$$T = \frac{2}{3 N k_B} \langle \mathrm{K} \rangle = \frac{2}{3 M N k_B} \sum_{\mu=1}^{M} \sum_{1}^{N} \frac{m_i (v_i^{\mu})^2}{2} \tag{6.19}$$

Here v_i^{μ} is the velocity of the i-th atom at the μ-th trajectory point of the simulation. Further M is the number of sampling points along the trajectory. The average kinetic energy was calculated using the above described expression (eq. 6.16) for calculation of averages of physical quantities.

The actual temperature of our system at a time *t* during the simulation is

$$T(t) = \frac{2}{3 N k_B} \sum_{1}^{N} \frac{m_i v_i^2}{2} \quad . \tag{6.20}$$

Calculation of pressure

The calculation of pressure is a more complicated task. First, we have to introduce the concept of virial originated from Clausius [4]. The virial function (W) of a molecule (system built from mass points) is defined as

$$W = \sum_{i=1} \mathrm{r}_i \mathrm{F}_i^{\mathrm{tot}} \quad , \tag{6.21}$$

where F_i^{tot} is the total force acting on the i-th atom at position r_i. It is easy to show that the time average of it is related to average of the kinetic energy as

$$\langle W \rangle = -2 \langle K \rangle \quad . \tag{6.22}$$

Using the formula (eq K-average) above we obtain

$$\langle W \rangle = -3 N k_B T \quad . \tag{6.23}$$

As a next step we separate the F_i to internal (interaction of the atoms) and external (interaction with the environment) forces

$$F_i = F_i^{\mathrm{int}} + F_i^{\mathrm{ext}} \quad \mathrm{i}=1,\ldots,\mathrm{N} \quad . \tag{6.24}$$

Assuming that the external forces are originated from the interaction of the atoms with a cuboid-like container, the contribution of the external forces to the time average of the virial is

$$\langle W^{ext} \rangle = \lim_{t\to\infty} \frac{1}{t} \int_0^t \sum_{i=1}^{N} r_i(\tau) \mathrm{F}_i^{ext}(\tau) d\tau \qquad (6.25)$$

$F_i^{ext}(\tau)$ is non-zero only for those particles which bounce to the wall of the container. For the sake of simplicity let us suppose that our container has a form of a rectangular prism with edges a_x,a_y,a_z. One of its corners is located at the origin of a coordinate system its edges are parallel to the axes. The forces at the six walls are perpendicular to the wall and directed to the inside of the container. Using this, the non-zero contributions to the virial average are

$$\langle W^{ext} \rangle = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{N} \left[\int_0^t -a_x F_{ix}^{ext}(\tau) d\tau + \int_0^t -a_y F_{iy}^{ext}(\tau) d\tau + \int_0^t -a_z F_{iz}^{ext}(\tau) d\tau \right] . \qquad (6.26)$$

The three terms can be rewritten using the definition of the pressure p,

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{N} \left[\int_0^t -a_x F_{ix}^{ext}(\tau) d\tau \right] = -a_x \cdot p a_y a_z \qquad (6.27)$$

where $a_y a_z$ is the surface area of that side of the rectangular prism which is parallel to the y-z coordinate plane. With this

$$\langle W^{ext} \rangle = -a_x \cdot p a_y a_z - a_y \cdot p a_y a_z - a_z \cdot p a_x a_y = -3pV \ , \qquad (6.28)$$

here $V=a_x a_y a_z$ is the volume of the container. Collecting our result

$$\left\langle \sum_{i=1}^{N} r_i F_i^{tot} \right\rangle - 3pV = -3Nk_B T \ . \qquad (6.29)$$

Or after rearrangement

$$pV = Nk_B T - \frac{1}{3} \left\langle \sum_{i=1}^{N} r_i F_i^{tot} \right\rangle \ . \qquad (6.30)$$

It is worth to mention that this equation reduces to the equation of state of the perfect gas if there is no interaction between the particles.

4. Environmental coupling: Thermostat, Barostat

To simulate thermodynamical ensembles we have to fix variables. Technically easy to tackle the problem if the volume (an extensive variable) is kept fixed in our simulation. We need simply to fix the geometrical parameters of the simulation box. However,it is more complicated to fix an intensive thermodynamical variable. As in reality, we need a thermostat or a barostat if we want to specify the temperature or the pressure of our system. The proper mathematical/computational representation is very important in MD because it inherently influences the probability distribution of the micro states obtained from our simulation and this way the calculated averages of thermodynamic quantities.

Following the categorization of G. Sutmann [5] we can distinguish four different methods for controlling a thermodynamic quantity *A*:

- *Differential control:* The value of *A* is fixed, no fluctuations are allowed
- *Proportional control:* The variables influencing the actual value of *A* are corrected towards the prescribed value of A in each simulation step. The 'speed' of correction is determined by a coupling constant which also

determined the fluctuation of *A* around its average value. This method simulates directly a system immersed in a 'bath' (*e.g.* thermostat or barostat).

- *Integral control*: In this case the environmental coupling is mimicked by adding extra degrees of freedom to the system which guaranties the prescribed value of *A.*
- *Stochastic control*: Certain degrees of freedom are modified stochastically to improve the control.

4.1. Temperature control

Temperature control

As we learned from statistical physics the temperature of molecular system containing N particles can be calculated from the average kinetic energy and the equipartition theorem: .

$$\langle K \rangle = \frac{3}{2} N k_B T \quad . \tag{6.31}$$

Through this connection the temperature and the particle velocities are interrelated as we have already seen at (6.20).

Differential control: simple velocity scaling

The simplest way to correct the temperature, if the calculated value T differs from the desired one, is to modify the velocity of the particles on a way which results in a proper average. The obvious modification is the velocity scaling (using $\lambda \mathbf{v}_i$, instead of $\mathbf{v}_i$ where λ is a scaling factor having normally a value close to one). The actual temperature of our system at a timepoint *t* during the simulation is

$$T(t) = \frac{2}{3Nk_b} \sum_1^N \frac{m_i v_i^2}{2} \quad . \tag{6.32}$$

After the velocity scaling the temperature will be equal to the desired T_0 value.

$$T_0 = \frac{2}{3Nk_b} \sum_1^N \frac{m_i \lambda^2 v_i^2}{2} \quad . \tag{6.33}$$

Subtracting it from we obtain

$$T_0 - T(t) = (\lambda^2 - 1) \frac{2}{3Nk_b} \sum_1^N m_i v_i^2 \;\Rightarrow\; \lambda = \sqrt{\frac{T}{T(t)}} \tag{6.34}$$

Here we used the (6.32) expression of T(t) . If this scaling is carried out at each step of the simulation the temperature will have a fixed value without any fluctuations (which is unphysical as in a canonical ensemble the kinetic energy has non-zero fluctuation) and the trajectory in the phase space will be discontinuous. A possible extension of this method is the so called Berendsen thermostat which simulates a weak coupling between the system and "heat bath" (energy reservoir).

Proportional control: Berendsen thermostat [6]

This thermostat is also a velocity rescale method, the rescaling occurs at every step, however, it is not complete, but damped. Mathematically the rate of temperature change is proportional to the temperature difference

$$\frac{dT}{dt} = \frac{1}{\tau}(T_0 - T) \quad , \tag{6.35}$$

where τ is the coupling parameter. This form leads to an exponential decay of the system towards T_0. From eq (6.35) using finite differences

$$\Delta T=\frac{\Delta t}{\tau}(T_0-T) \tag{6.36}$$

With the (6.32) and (6.33) equations for the temperatures T and T_0 we can calculate the scaling factor of velocities that leads to the temperature difference ΔT in a time of τ

$$(\lambda^2-1)\frac{2}{3Nk_B}\sum_1^N\frac{m_i v_i^2}{2}=\frac{\Delta t}{\tau}(T_0-T),$$
$$\lambda=\left(1+\frac{\Delta t}{\tau}\left(\frac{T_0}{T}-1\right)\right)^{1/2} \tag{6.37}$$

The Berendsen thermostat does not generate proper canonical ensemble because it suppresses the fluctuations of the kinetic energy. Fortunately, with the introduction of a stochastic term this problem is solvable [7].

Integral control: Nose-Hoover thermostat [8,9]

In this case the thermal coupling is mimicked by adding extra degrees of freedom to the system which guaranties the prescribed value of temperature. The main advantage of this method is that the temperature control is not an external procedure but included in the equation of motion. A new virtual atom with "mass" M and a "coordinate" of Q is introduced. The equation of motion of our extended system is:

$$m_i\boldsymbol{a}_i=F_i-\frac{\partial Q}{\partial t}m_i\boldsymbol{v}_i, \qquad i=1,\ldots,N$$
$$M\frac{\partial^2 Q}{\partial t^2}=\sum_{i=1}^{N}m_i v_i^2-gk_BT \quad . \tag{6.38}$$

Here *g* is the degrees of freedom of the system, k_B is the Boltzmann constant, T is the absolute temperature and F_i is the internal force acting on atom i. With the proper selection of M the virtual particle ensures the temperature control of the system. The "force" introduced here at the right hand side of the last equation is small if the kinetic energy is close to that given by the equipartition theorem and large if it is far from it. This method is called Nosé-Hoover thermostat [8,9].

Stochastic control: Langevin thermostat

This method has its origins in the Langevin stochastic differential equation of motion which describes the motion of atoms due to a thermal agitation of a heat bath

$$m_i a_i=F_i+\gamma m_i v_i+F_i^{rand}, \qquad i=1,\ldots,N \tag{6.39}$$

where γ is the friction coefficient and F_i^{rand} is a rapidly varying random force (with zero average) due to the coupling of the system to the many degrees of freedom of its environment. This method through the last two terms in the equation of motion simulates the collision of the system with the particles of the environment.

4.2. Pressure control

The fundamental strategies of pressure control are tight analogs of the temperature control methods we have seen above. Here we mention only two important methods, the Berendsen and Parrinello-Rahman pressure controls.

Proportional control: Berendsen barostat

In the case of temperature control we have introduced a scaling procedure for the velocities, which corrected the temperature towards the desired. There we used the connection between the temperature and the kinetic energy of the system. In the case of pressure we can rescale the positions and this way system volume on the same way.

Mathematically the rate of pressure change is proportional to the pressure difference

$$\Delta P = \frac{\Delta t}{\tau_p}(P_0 - P) \ , \tag{6.40}$$

where P_0 is the pressure we want to keep, τ_{p0} is the coupling parameter. This form leads to an exponential decay of the system towards P_0. To make corrections for reaching the pressure P_0 we need to rescale all of the coordinates of the system at every step of the integration by a factor of

$$\mu = \left[1 - \frac{\Delta t}{\tau_p}(P_0 - P)\right]^{1/3} \tag{6.41}$$

This gives the case of isotropic compression, which can also be generalized for anisotropic case.

Integral control: Parrinello-Rahman barostat

The *Parrinello-Rahman*barostat is analogous to the Nosé-Hover thermostat where the extra degree of freedom was used to simulate the effect of weak coupling of the system to a heat bath. In this case the extra degree of freedom mimics a "piston" which corrects the pressure toward its desired value.

5. Constraints

A biomolecule can be characterized in many cases by an extremely large number of geometrical parameters like bond length, bond angles or torsion angles etc. The parameters are not equally important in the simulation of characteristic features of a molecule. Some of them can be frozen without significant influence on the result obtained from the simulation. It can also happen that we want to investigate well defined conformations of a molecule (L or D conformation of an amino acid) that would change in a calculation (*e.g.* at higher temperatures) without fixing them. There are other practical reasons to fix a bond length. If the bond is described by a deep and narrow potential valley, its oscillations are to fast and requires smaller time step in the numerical integration of the equation of motion.

This type of constraint can be given in general as

$$\sigma_k(r_1, r_2, \ldots, r_N) = 0, \qquad k = 1, .., N_c \tag{6.42}$$

Where N_C is the number of constraints. In the simplest case it has the form of

$$\sigma_k(r_1, r_2, \ldots, r_N) = |r_{a_k} - r_{b_k}|^2 = 0, \qquad k = 1, .., N_c \tag{6.43}$$

a_k and b_k are equal to atomic indices of the constrained atoms. This type of constraint is called *holonomic* in classical mechanics and can be incorporated to the Newton's equation in the form of an additional force

$$F_i + G_i = m_i a_i, \qquad i = 1, .., N, \tag{6.44}$$

where

$$G_i = -\sum_k \lambda_k \nabla_i \sigma_k, \qquad k = 1, .., N_c, \tag{6.45}$$

λ is the Lagrange multiplier, G_i is the constraint force, well known from basics of classical mechanics. This is the force, for example, acting on a body sliding down on a slope, perpendicular to the slope and ensures that the body remains on the surface.

Having constraints, the equation of motions contains 3N unknown coordinates and N_c undetermined Lagrange multipliers. To the solution we have 3N equation from Newton's second law and N_c equations of the constraints. This ensures the solvability of the problem.

There are several methods of incorporation of the constraint into the numerical integration schemes of the equation of motion. We mention here the SHAKE method where the numerical integration step is made without any constraint first obtaining a new set of atomic coordinates. The coordinates are modified in the second step using an iterative method to satisfy the constraints (for the details see [10,11]). The LINCS algorithm [10,12] (implemented in *e.g.* GROMACS package) works similarly but it uses a non-iterative single step procedure to

restore the constrained distances after an unconstrained step. It is faster and more stable than SHAKE but it can only be used with bond length constraints and isolated angle constraints, such as the proton angle in X-OH.

6. Advanced MD-based methods: Simulated annealing, REMD

For small molecules the energy hyper-surfaces are relatively simple. Their main features (minima, transition states *etc.*) can be determined easily using e.g., systematic conformational search methods. In this case we use local optimization which provides mostly the closest minimum to the initial geometry. If we start this search from a well selected set of initial states we have a good chance to find every single minimum. However, for large biomolecules these methods cannot be carried out on the same way. On the other hand, the determination of a single energy minimum has not got the same importance as it has for small systems. Normally, a biomolecule has several different energy minima related to a certain functional form of it. In a living organism (non-zero absolute temperature) most of these minima are realized (with different probability determined by the energy difference appearing in the Boltzmann distribution). Often a macromolecule can have several from these kinds of minima sets (different states) which are separated by energy barriers of different heights. If we want to determine the functionally different structures of macromolecules, we have to find these sets of energy minima. The methods used for this purpose are closely related to *global optimization techniques* of numerical mathematics.

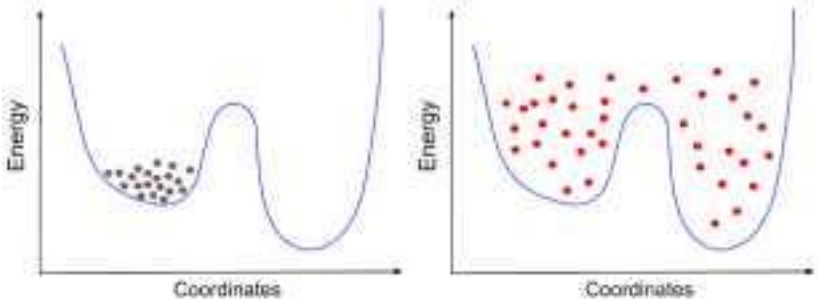


Figure 6.6. Schematic representation of an atomic system trapped in a local minimum at low temperature (left panel) and its escape from there at higher temperature (right panel). Blue curve represent the potential energy surface of the system while gray and red dots show possible total energies of the system at low and high temperatures, respectively.

The *simulated annealing* method **[13]** is the first to mention here. The name of this method has its origins in metallurgy. In this process a material is first heated and than slowly cooled to improve its crystal structure reducing the defects in it. This method is used in steal production resulting in improved strength and durability of the product.

The numerical method mimics this procedure. We first heat our system to a “high temperature” and let it equilibrate there. This step ensures that the kinetic energy “fills up” the potential energy valleys (*Figure 6.6.*), which makes it possible for the system to escape from being trapped there. As a final step, the system is cooled to a low temperature slowly. Theoretically, this method should provide a global minimum of the potential surface related to our biomolecule. But in practice we obtain only a low energy conformation. The result depends on the protocol used: the speed and functional form and final temperature of heating, the length of equilibration and the speed and functional form (linear, stepwise or exponential) of cooling. Often the simulated annealing cycle is repeated several times and the final structures are used as representatives of the low energy conformers.

The other method we mention here is the *replica exchange molecular dynamics* (REMD) [14] which have became more and more popular during the last decade. This method is planned to provide proper statistical ensembles at different temperatures simultaneously. But it is also able to avoid being trapped in a certain minimum, which may happen in the case of a single, low temperature dynamics. This method is often applied for the conformational analisys of flexible molecules like olygopeptides (say peptapetides). The applicability is strongly limited by system size because the available conformers grows rapidly with rotable bonds.

The method consists of parallel simulations of the same system (called replicas) at different temperatures. After a constant temperature simulation period, the temperatures of replicas (i-th and j-th) are exchanged with the probability of:

$$P_{ij} = min\left(1, \exp\left[\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)(U_i - U_j)\right]\right), \tag{6.46}$$

where T_i and T_j are temperatures of two replicas; U_i and U_j are the corresponding potential energies. In the practical implementations exchanges are restricted to the neighbouring replicas. The temperature selection is crucial for the proper working of the method. If the temperatures are to far from each other the exchange probability is too low and the method will become simple simultaneous MD-s at different temperatures. For the prediction of a proper temperature set see Ref . [15] and a "REMD calculator" at [16]. REMD is often used for the generation of conformational ensembles for small or middle-size biomolecules (like polypeptides or small proteins) to investigate the influence of the temperature rising on their structural stability and other physical quanties (like helicity or solvent accessible surface area). The influence of the environment (i.e. cosolutes) and structural alterations (like point mutation of a protein sequence) on these quantities can also be elucidated using this method.

7. Summary

In this chapter an outlook of the basic concepts of molecular dynamics was given. The problems of the selection of a model system were discussed, including the fundamentals of periodic boundary condition method, well known from solid state physics. Three simple numerical integration procedures were also detailed for the solution of Newton's equation motion of a molecular system.

The largest field of applications of MD is the calculation of the averages of physical (*e.g.* thermodynamical or structural) quantities. The statistical mechanical backgrounds of the related procedures were described from the basic concepts (like ensemble or calculation of averages) to the technical details. We discussed the methods applicable in constant temperature (thermostats) and constant pressure (barostats) calculations as well as the possibilities of the inclusion geometrical constraints. Finally we discussed two popular methods of advanced MD. The first was the simulated annealing which is used for the sampling of low energy conformers of molecule. The second was the replica exchange molecular dynamics planned to provide proper statistical ensembles for an extended system at different temperatures simultaneously.

The field of molecular dynamics is developed very intensively due to the spectacular improvement of computers, recently. The interested reader may find further information on the new methods as well as on their implementations on different computer architectures in the section of "Further readings". Some comprehensive MD books and the availability of the most popular MD codes are listed there, as well.

8. References

1. L.Verlet, „Computer Experiments on Classical Fluids. I. Thermodynamical Properties of Lennard−Jones Molecules". *Physical Review***159**: 98–103 (1967); http://en.wikipedia.org/wiki/Verlet_integration

2. http://en.wikipedia.org/wiki/Leapfrog_integration

3. W.C. Swope, H.C. Andersen, P.H. Berens, K.R. Wilson, „A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters", *The Journal of Chemical Physics***76** (1): 648(Appendix)(1982); http://en.wikipedia.org/wiki/Verlet_integration#Velocity_Verlet

4. R.J.E. Clausius,. "On a Mechanical Theorem Applicable to Heat", Philosophical Magazine, Ser. 4 40: 122–127 (1870).

5. G. Sutmann, „Quantum Simulations of Complex Many-Body Systems:From Theory to Algorithms, Lecture Notes", J. Grotendorst, D. Marx, A. Muramatsu (Eds.),John von Neumann Institute for Computing, Jülich, NIC Series, Vol. 10, ISBN 3-00-009057-6, pp. 211-254 (2002); http://www2.fz-juelich.de/nic-series/volume10/sutmann.pdf

6. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak,. "Molecular-Dynamics with Coupling to an External Bath", Journal of Chemical Physics **81** (8): 3684–3690 (1984).

7. G. Bussi, D. Donadio, M. Parrinello, „Canonical sampling through velocity rescaling", J. Chem. Phys. 126 014101 (2007).

8. S. Nosé, „A unified formulation of the constant temperature molecular-dynamics methods", Journal of chemical physics **81** (1): 511–519 (1984); http://en.wikipedia.org/wiki/Nos%C3%A9%E2%80%93Hoover_thermostat

9. W. G.Hoover, „Canonical dynamics: Equilibrium phase-space distributions", Phys. Rev. A **31** (3): 1695–1697 (Mar 1985).

10. D.C. Rappaport, *The Art of Molecular Dynamics Simulation, 2nd ed.,* Cambridge University Pres, Cambridge pp. 264, 2004.

11. J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, "Numerical integration of the cartesian equations of motion for a system with constraints: Molecular dynamics of n-alkanes", J. Comp. Phys. 23, 327 (1977).

12. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations", J. Comp. Chem. 18:1463–1472 (1997).

13. S. Kirkpatrick, C.D. Gelatt, M. P. Vecchi, „Optimization by Simulated Annealing", Science **220** (4598): 671–680 (1983); http://en.wikipedia.org/wiki/Simulated_annealing

14. Y. Sugita and Y. Okamoto Replica-exchange molecular dynamics method for protein folding, Chemical Physics Letters **314**: 141–151 (1999); http://en.wikipedia.org/wiki/Parallel_tempering#cite_note-4

15. A. Patriksson and D. van der Spoel, „A temperature predictor for parallel tempering simulations" Phys. Chem. Chem. Phys., 10, 2073-2077 (2008).

16. http://folding.bmc.uu.se/remd/

9. Further Readings

A comprehensive discussion of the methods of molecular simulations can be found in

- R. Leach, *Molecular Modelling, Principles and Applications*, 2nd ed, Prentice Halls, pp. 603-608, Pearson Education Limited, 2001.
- D.C. Rappaport, *The Art of Molecular Dynamics Simulation* 2nd ed., Cambridge University Pres, Cambridge 2004.
- T. Schlick, *Molecular Modeling and Simulation*. Springer, 2002.

The most popular MD program packages are listed below together with the location of their web pages and tutorials.

Package	***URL***	***Tutorial***	***Free?***
AMBER	ambermd.org	ambermd.org/tutorials	-
CHARMM	www.charmm.org/	www.charmmtutorial.org/	-
GROMACS	www.gromacs.org	www.gromacs.org/Documentation/Tutorial	+
NAMD	www.ks.uiuc.edu/Research/namd	www.ks.uiuc.edu/Training/Tutorials	+
DESMOND	*www.deshawresearch.com/resources_desmond.html*	www.deshawresearch.com/Desmond_Tutorial_0.6.1.pdf	+

10. Questions

1. What is the periodic boundary condition?
2. How are the non-bonding forces truncated in a periodic boundary condition simulation?
3. Derive the Verlet integration method applied for the numerical solution of Newton's equation.
4. What is the connection between the Verlet and leapfrog integration methods?
5. What are the advantages of the velocity Verlet integrator?
6. Characterize the statistical ensembles most often used in simulations.
7. How are the average values of physical quantities calculated from the data, collected during a simulation?
8. How is the average pressure calculated?
9. What is the Berendsen thermostat?
10. Explain the theory behind the LINCS and SHAKE methods.
11. What is the simulated annealing method?
12. How does the replica exchange molecular dynamics work?

11. Glossary

- *Cluster:* Collection of a small amount of interacting atoms or molecules.
- *P eriodic boundary condition*s (PBC): A method that helps to solve the problem of overemphasized surface effects of cluster simulations. The 3D space is covered by identical copies of a repeat unit.
- *R eference cell*: Repeatunit in a PBC calculation.
- *Micro state:* Every possible combination of the position and momentum vectors of a system.
- *Macro state:* A state of a system characterized by fixed values of thermodynamical variables.
- *Phase space:* Collection of all possible microstates.
- *Ensemble:* A probability distribution of the micro states.
- *Trajectory:* A curve in the phase space that consist of those points which were reached during the time evolution of the system.
- *Ergodicity:* A system is *ergodic*, if a single copy of the system will go through all of its microstates, if we follow its evolution (*trajectory*) in the state space for an appropriately long time.
- *Thermostat:* Here it denotes a mathematical construction, which keeps the system temperature at a desired value in average during the simulation.
- *Barostat:* Here it denotes a mathematical construction, which keeps the system pressure at a desired value in average during the simulation.
- *Simulated annealing (SA):* An MD based global optimization procedure. During SA the system is heated and cooled subsequently.
- *Replica exchange molecular dynamics (REMD):* The method consists of parallel MD simulations of the same system (called replicas) at different temperatures. After a constant temperature simulation period, the temperatures of replicas are exchanged using a Monte Carlo-like criterion.

Chapter 7. Prediction of Protein Structures and a Part of the Protein Structure

(Tamás Körtvélyesi)

Keywords: protein secondary structures, missing 3D structures, structures of loops, ab initio protein structure prediction

What is described here? The 3D structures from the results of XED or NMR are insufficient, but the primary structure (sequence) is known. For medelling the structurein MD or in docking procedure, the best if we know the whole structure. On the basis of known sequences and the 3D structures of these sequences, supposed that the sequence and structure similar in different proteins, the missing part can be built up.

What is it used for? To repair the 3D structure of proteins with missing 3D residues for modelling by MD and docking proteins and small drug like molecules.

What is needed? The basic knowledge of the structure of peptides, proteins, molecular mechanics, pK calculations of the side-chains are important.

1. Introduction

The knowledge of the protein structure is important in the molecular modelling of different reactions proteins (association with proteins and/or small molecules). In some cases only the sequence is known and no other information is available ont he 3D structure. In this case *ab initio* structure prediction is necessarry. This method is under developement and the results are a lot of times not acceptable.

Figure 7.1. Global fitting

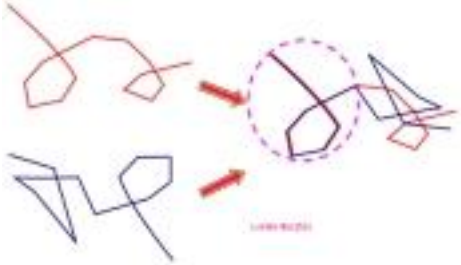


Figure 7.2. Local fitting

2. *Ab initio* Protein Structure

If the protein structure is known by sequence, but the 3D structures are not known, sometime it is important to predict the 3D structures for molecular modelling.

3. Threading

The basic principle of threading is that an unknown amino acid sequence is fitted into existing known 3D structure and after the fitted sequence is folded into the structure is evaluated. It means, that the side chains are not known, only the backbone structure.

4. Homology Modelling and Loop Prediction

4.1. Sequence analysis, Pairwise Alignment and multiple sequence alignment

In many cases the structure of proteins is known only in parts in XRD and NMR experiments. Though, the sequence is known we would like to know the 3D structure. One approach to perform pairwise alignment with a protein which have in some positions the same sequence and its 3D structure is known. Other, better method to perform the alignment with more than two proteins. This process is the multiple sequence allignment.

Figure 7.3. Sequences of BSA and HSA (the yellow background is for the conservative aminoacids)

In the evolution in the proteins with both evalutanary and structural similarity of different species mutations occurred. Some residues were changed with the same hydrophilic or hydrophobic ones. In some cases the residues were different in its character. In multiple alignments where the sequence is similar, the structure superimposable to each other. Manually, the mutiple alignment is not very precise and time consuming. Generally used algorithms are the Needlham-Wunsch and the Smith-Waterman algorithms. They use pairwise alignment from the starting point. The result is aligned with the next sequence and so on. The Greedy algorithms share the problems small pieces and not as a whole problem. One of the generally used program is the ClustelW. (see http://en.wikipedia.org/wiki/Homology_modeling)

Table 7.1. Programs and servers for homology modelling (the sources see the Table 7.2)

Software	Method	Notes
3D-JIGSAW	Fragment assembly	Automatic webserver
CABS	Reduced modelling	Downloadable software
CHPModel	Fragment assembly	Automatic webserver
EsyPred3D	Template searching, alignment, 3D modelling	Automatic webserver
GeneSilico	Consenzus template searching, fragment assembly	Webserver
Geno3D	Segment matching	Automatic webserver
Hhpred	Template searching, alignment, 3D modelling	Automatic webserver
LIBRA I	Light Balance for Remote Analogous proteins	Webs erver
MODELLER	Segment matching	Downloadable software
ROSETTA	Rosetta homology modelling and ab initio fragment assembly	Webserver
SWISS MODEL	Local similarity, fragment assembly	Automatic webserver
TIP-STRUCTFAST	Automatic comparision modelling	Webserver

Software	Method	Notes
WHAT-IF	Position specific rotamers	Webserver

4.2. Steps of modelling

Modelling includes four steps: (i) choose of the template, (ii) target-template fitting by using a score function, (iii) build up modells and (iv) evaluation of the modells. The first two steps sometimes handled together.

4.3. Choose of the template (i), target-template fitting by using a score function (ii)

The procedure is the sequence fitting (FASTA and BLAST) on the basis of the character of the aminoacids: the same aminoacids in different in the same proteins are conservative residues, hydrophobic aminoacids and negative/positive charged residues. The fitting can be performed by using score (penalty) functions. The simple fitting uses alignment or comparision and fitting more sequences by multiple alignment.

Table 7.2. Softwares and their source in the internet

Usage	Program	Source in the internet
Sequences	UniProt databasis	http://www.uniprot.org/
Sequence analysis	CLC Sequence Viewer 6.3	http://www.clcbio.com/index.php?id=28
Homológy modelling	3DjigsawEsyPred3D	http://bmm.cancerresearchuk.org/~3djigsaw/
	Lomets	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
	SwissModel	http://zhang.bioinformatics.ku.edu/LOMETS/
	Geno3D	http://swissmodel.expasy.org/
	CBS	http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html
		http://www.cbs.dtu.dk/services/CPHmodels/
Protein structure, comparision, evaluation of modells	Dali	http://ekhidna.biocenter.helsinki.fi/dali_lite/start
	Matras	http://biunit.aist-nara.ac.jp/matras/matras_pair.html
	SuperPose	http://wishart.biology.ualberta.ca/SuperPose/
	CATH	http://www.cathdb.info/cgi-bin/SsapServer.pl
	CeCalculator	http://cl.sdsc.edu/ce/ce_align.html
Optimization , molecule mechanics	TINKER	http://dasher.wustl.edu/tinker/
Molekuladynamics	Gromacs	http://www.gromacs.org/
pKa prediction	Vega ZZ (Propka2.0)	http://www.vegazz.net/

Usage	Program	Source in the internet
Ramachandran plot	VMD	http://www.ks.uiuc.edu/Research/vmd/
Molecule graphics	Molegro	http://www.molegro.com/
	PyMOL	http://www.pymol.org/
	ICM Browser	http://www.molsoft.com

4.4. Choose of the template (i), target-template fitting by using a score function (ii)

The procedure is the sequence fitting (FASTA and BLAST) on the basis of the character of the aminoacids: the same aminoacids in different in the same proteins are conservative residues, hydrophobic aminoacids and negative/positive charged residues. The fitting can be performed by using score (penalty) functions. The simple fitting uses alignment or comparision and fitting more sequences by multiple alignment.

4.5. Generation of modells

The generation of the modells has 3 main methods [1].

1. Fragment assembly

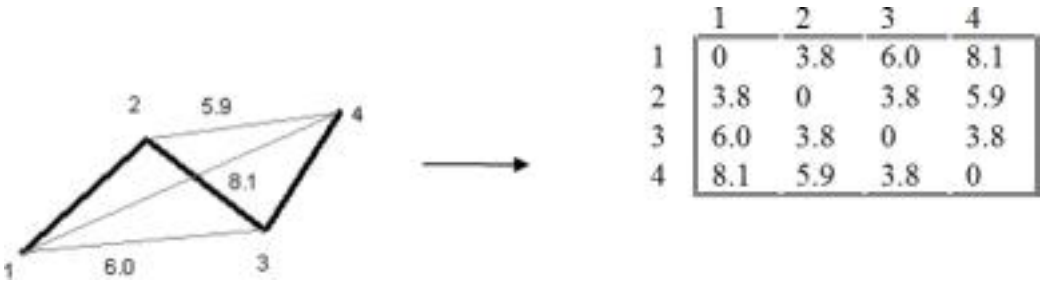
   The common fragments are shifted to build up the missing part of the protein. The 3D structures of the fragments are coupled to each other. Modelling structures of the loop regions are difficult.

2. Homology modelling based on constraints

   The procedure does not share conservative and movable part of the missing protein structure. Comparision of sequences by geometry (torsion angle, distance between C_α atoms, torsion angles in side-chains, constraint of the backbone length of peptide/protein. The geometrical criteria is fitted.

3. Homology modelling based on segment metching

   The target is shared to small segments and its templates are fitted from the Protein Data Bank. The distance of C_α atoms are compared and predict the strain int he templates and int he predicted structure based on van der Waals radii. (see the distance matrix in *Figure 7.4* and Eq. 7.1).



	1	2	3	4
1	0	3.8	6.0	8.1
2	3.8	0	3.8	5.9
3	6.0	3.8	0	3.8
4	8.1	5.9	3.8	0

Figure 7.4. Distance matrix

RMSD calculation

$$RMSD(A,B)=\sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i,[b_i])^2} \qquad (7.1)$$

N is the number of atoms, $d(a_i, b_i)$ are the distances between a and b atoms. The superposition of rigid protein structure gives also information ont he goodness of homology modelling The minimization of ε gives the error (Eq. 7.2):

$$\varepsilon = min \sum_{i=1}^{N} \left| \left| T(a_i) - b_i \right| \right|^2 \qquad (7.2)$$

Two sets of points are given: A = (a_1, a_2,… a_n) and B = (b_1, b_2,… b_n) in 3D. The optimal part set of A(P) and B(Q) (absolute value) are searched. We try to find the optimal rigid body transformations (G_{opt}) between A(P) and B(Q) sets which gives the minimal distances (D) by the rigid body transformations..

To compare the structures there are some softwares: Dali [2], CE CALCULATOR [3], MATRAS [4], SSAP (CATH) [5], TIP-STRUCTFAST [6], SuperPose [7], VAST [8.] (see Table 7.4.)

If the sequence identity is less than 30%, our modells are unreliable, between 30-60% the modell is reliable, buti t has unreliable regions, at larger than 60% our modell is reliable. We have to be careful if homology modelling is near the binding pocket. Binding pocket is sensitive enough, we have to „positively believe" in our results and handle with critics.

Table 7.3. Some softwares to compare the protein structures

Name of the software	Input	Output
Dali [2]	pdb file/pdb code	structure fitting and alignment
Cath (SSAP) [5]	pdb file/pdb code	structure fitting and pairwise alignment
Ce calculator [3]	pdb file/pdb code	pdb file, structure alignment (Rasmol)
Matras [4]	pdb file/pdb code	structure fitting and pairwise alignment, pairwise 3D alignment
SuperPose [7]	pdb file/pdb code	sequence fitting and alignment (PDB), RMSD

The results of the homology modelling with different servers of BSA (Bovine Serum Albumine) from HSA (Human Serum Albumine) structure) are described in *Figure 7.5.-Figure 7.10.* As it can be seen the structures are similar after homology modelling but we can find differences int he positions of secondary structure. Table 7.5 includes the frequencies of aminoacids in BSA and HSA for modelling homology modelling. As it can be seen, the frequency of aminoacids is very similar between BSA and HSA.

Table 7.4. The frequency and the number of aminoacids in BSA and HSA

Aminosav	**HSA**	**BSA**	**Aminosav**	**HSA**	**BSA**
Alanin (A)	0,103	0,079	Alanin (A)	63	48
Cisztein (C)	0,057	0,058	Cisztein (C)	35	35
Aszparginsav (D)	0,069	0,066	Aszparginsav (D)	36	40
Glutaminsav (E)	0,102	0,097	Glutaminsav (E)	62	59
Fenilalanin (F)	0,057	0,049	Fenilalanin (F)	35	30
Glicin (G)	0,021	0,028	Glicin (G)	13	17
Hisztidin (H)	0,026	0,028	Hisztidin (H)	16	17
Izolucein (I)	0,015	0,025	Izolucein (I)	9	15
Lizin (K)	0,099	0,099	Lizin (K)	60	60

Aminosav	**HSA**	**BSA**	**Aminosav**	**HSA**	**BSA**
Leucin (L)	0,105	0,107	Leucin (L)	64	65
Metionin (M)	0,011	0,008	Metionin (M)	7	5
Aszpargin (N)	0,028	0,023	Aszpargin (N)	17	14
Prolin (P)	0,039	0,045	Prolin (P)	24	28
Glutamin (G)	0,033	0,033	Glutamin (G)	20	20
Arginin (R)	0,044	0,043	Arginin (R)	27	26
Szerin (S)	0,046	0,053	Szerin (S)	28	32
Treonin (T)	0,048	0,056	Treonin (T)	29	34
Valanin (V)	0,071	0,063	Valanin (V)	43	38
Triptofán (W)	0,003	0,005	Triptofán (W)	2	3
Tirozin (Y)	0,031	,0035	Tirozin (Y)	19	21

Figure 7.5. 3D-JIGSAW (web server) BSA predicted structure. The method is fragment based, the whole sequence of BSA is necessarry.

On the basis of the frequency of aminoacids in BSA and HSA, it can be seen that they are very similar.

Figure 7.6. EsyPred3D (web server): BSA predicted structure. The method is fragment based, the whole sequence of BSA is necessarry. The applied structure was 1AO6 (HSA). The template-target identity 72,6%.

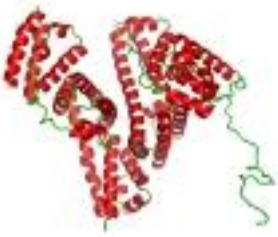


Figure 7.7. LOMETS (web server): BSA predicted structure. The method is fragment based, the whole sequence of BSA is necessarry. The applied structure was 1n5u (HSA). The template-target identity 72,6%.

Figure 7.8. Swiss-Prot (web server): BSA predicted structure.

Figure 7.9. Gen3D, homology modelling with constraints.

Figure 7.10. CBS (web server): BSA predicted structure.

The real XRD structure of BSA was published: PDB Id.:3V03. The molecule is important in the bionano experiment to hydrophylize (capsulate) hydrophobic drug molecules in drug delivery. A checking homology modelling was applid for HSA from the obtained BSA model. Its goodness (RMSD) can be seen in Table 7.6.

Table 7.5. RMSD values calculated by VMD [9] (the numbers in parenthesis are the number of residues considered int he calculations)

	HSA	HSA	CBS	ESYPRED	GENO	LOMETS	SWISS
HSA		24,77(480)	24,77(480)	0,4423(570)	5,626(420)	24,74(475)	5,788(495)
3DJIGSAW	24,77(480)		24,45(545)	25,18(565)	24,93(545)	0,753(485)	24,69(565)
CBS	24,77(480)	24,45(545)		4,136(445)	4,451(545)	24,44(485)	0,7926(565)
ESYPRED	0,4423(570)	25,18(565)	4,136(445)		6,037(515)	24,68(485)	5,989(565)
GENO	5,626(420)	24,93(545)	4,451(545)	6,037(515)		24,74(485)	4,338(565)
LOMETS	24,74(475)	0,753(485)	24,44(485)	24,68(485)	24,74(485)		24,60(485)
SWISS	5,788(495)	24,69(565)	0,7926(565)	5,989(565)	4,338(565)	24,60(485)	

To refine the structure, it is important to optimize the structure. The pKa of the side chains were calculated. The structures were optimized by TINKER/AMBER99 and AMBER99/GBSA (see Chapter 1). The result compared with the unoptimized structures can be seen in *Figure 7.11*.

The structure is acceptable after optimization by real empirical folding force field foldX [10], which is an excellent method after homology modelling.

A good choice in homology modelling and analysis of proteins is the CLCBio [11]. The structures after the generation of the structures must be optimized. The strains can be decreased by optimization and after this procedure MD calculations (e.g. gromacs) are necessarry [12,13].

Figure 7.11. Optimized geometries of the BSA structures ontained by homology modelling (TINKER/AMBER99 and TINKER/AMBER99/GBSA)

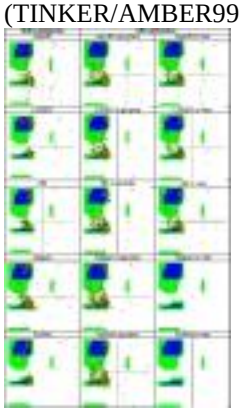


Figure 7.12. Ramachandran plots of BSA before optimization and after optimization (the optimization of the empty plots were not successful).

For the initial structure of the proteins it is suggested to perform MD calculations of the structures obtained by XRD. In the end of the homology modelling or loop prediction to reduce the strain also MD simulation is suggested. Fort he optimal initial structure of side chains can be obtained by SCWRL4 method [14].

Docking with ligands can help to validate the structure of the proteins obtained by homology modelling. The ligand-protein complex structure have to be known.

5. Summary

The missing 3D structures in proteins can be predicted by homology modeling. The *ab initio* structure prediction is not really reliable, our knowledge is lacking. Different strategies are supported by the homology modelling methods. There are some methods to check the goodness of the models.

6. References

1. D Baker, A Sali, Protein Structure Prediction and Structural Genomics , Science 294, 93-96 (2001).

2. http://ekhidna.biocenter.helsinki.fi/dali_lite/start

3. http://www.cathdb.info/cgi-bin/SsapServer.pl

4. http://cl.sdsc.edu/ce/ce_align.html

5. http://biunit.aist-nara.ac.jp/matras/matras_pair.html

6. http://amazon-tip64.eidogen-sertanty.com/Login.po

7. http://wishart.biology.ualberta.ca/SuperPose/

8. http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

9. W. Humphrey, A. Dalke, K. Schulten, **VMD - Visual Molecular Dynamics.** J. of Mol.

   Graph., 14:33-38(1996).

10. J. W. Schymkowitz, F. Rousseau, I. C. Martins, J. Ferkinghoff-Borg, F. Stricher, L. Serrano L., Prediction of water and metal binding sites and their affinities by using the Fold-X force field.Proc Natl Acad Sci U S A,102(29):10147-52(2005).

11. http://www.clcbio.com/index.php?id=28

12. http://www.gromacs.org

13. http://research.ozreef.org/GROMACS_MD_Flowchart.pdf

14. http://dunbrack.fccc.edu/scwrl4/SCWRL4.php

7. Further Readings

1. Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins. 2nd Ed., Ed. A. D. Baxevanis, B.F. F. Quelette,Wiley Interscience John Wiley & Sons, Inc. Publication, 2001.

2. C. Gibas, P. Jambeck, Developing Bioinformatics Computer Skills, O'Reilly & Associates, Inc. 2001.

3. Structural Bioinformatics, Ed. by P. E. Bourne, H. Weissig, Wiley-Liss, A John Wiley & Sons Publication, 2003.

8. Questions

1. What is the *ab initio* prediction of protein structure?

2. What is the homology modeling and loop prediction?

3. What is the procedure of homology modeling?

4. What methods are used in the model generation in homology modelling?

5. What methods are necessarry after the prediction of the structure in homology modelling?

6. What methods are known to support the goodness of homology modelling?

7. What is threading procedure?

9. Glossary

Ab initio **protein structure prediction** The known sequence without 3D structure demand to predict the 3D structure for modeling. This method can be only a suggestion on the structure, It can contain failures.

Homology modeling The missing 3D structures can be predicted considering known parts of structures with different methods.

Chapter 8. Protein-protein and Protein-ligand Binding. Docking methods

(Tamás Körtvélyesi)

Keywords: Docking of protein-protein, docking of protein-ligands, drug-like molecules, scoring, rescoring, discovering of binding sites

What is described here? The docking protocols of protein to protein and protein to ligands (drug like molecules) algorithms are described in this chapter.

What is it used for? The procedure is basic in the computational assisted drug design (CADD). We can obtain informations on the binding mode(s) and binding free energy which latter value can be compared the results of the biological experiments. An acceptable method to predict drug-like molecules which can inhibit the reactions in the binding site of proteins. The method with modifications is suitable for the prediction of binding site on the protein to determine the structure-function relationship if the binding site is unknown..

What is needed? The knowledge of the inter- and intramolecular interactions between molecules and in the molecules, respectively, is basic for the calculations in docking procedure. The structure of biomolecules and the calculation of the potential functions in the (Coulomb and van der Waals) interactions are also important.

1. Introduction

The protein to protein and protein to ligand docking is one of the most important computational procedure to predict the protein association and the association of protein and ligand (drug-like) molecules In the association the shape complementarity and the electrostatic complementarity determine the possible structures (Koshland [1]). The strategies and methods in modeling are different. Generally, a lot of structures are generated with the shape complementarity and the electrostatic complementarity. The structures are ranked on the basis of the values in score function. In the next step more sophisticated methods are available for rescoring to find the acceptable structures.

2. Protein-protein Docking

The idea of the protein-protein docking is based on the shape complementary calculated by a geometric recognition algorithm in 3D with Fourier transformation which was developed by Katchalski-Katsir et al. [2]. The two rigid molecules are denoted as a and b. woth N x N x N dimensional grids (see Fig. 8. 1. in 2D). The discrete functions are in Eq. (8.1) and Eq. (8.2).

$a_{l,m,m}$=1, inside the molecule and 0 outside the molecule (8.1)

$b_{l,m,m}$=1, inside the molecule and 0 outside the molecule (8.2)

where l, m, n are the indices in the 3D grid.

The difference in the surface and the interior of the two molecules can be defined by Eq. (8.3) and Eq. (8.4).

$a'_{l,m,n}$=1, on the surface of the molecule, ρ inside the molecule an 0 outside the molecule. (8.3)

$b'_{l,m,n}$=1, on the surface of the molecule, δ inside the molecule an 0 outside the molecule. (8.4)

Matching the surface, a correlation function is used. The correlation functions are transformed by discrete Fouriertransformation (DFT) (see Figure 8.1). This shape complementary calculation is the basic almost all of the protein-protein docking. The main differences are in the score function which can be empirical or based on the van der Waals and electrostatic interactions.

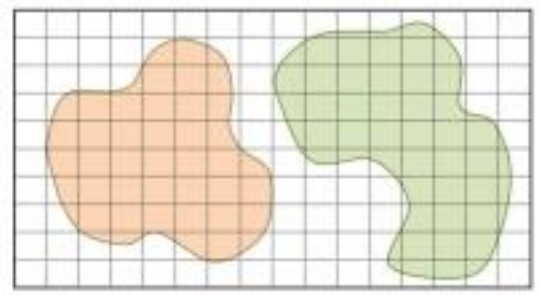


Figure 8.1. A model of two proteins in a 2D grid

It is suggested first to perform molecular dynamics calculation of the proteins to solvate the XRD results in explicit water molecules. The original idea of shape complementary calculation and newly developed score function can be found in MOLFIT [3]. Hydrophobic score function is built in GRAMM [4] and GRAMM server [5]. HEX [6] has new mathematical procedures and it can use GPU/CUDA to make the procedure much more faster. On the calculation of the binding free energies by rescoring, see the session of **Rescoring**.

3. Protein-Small Molecule Docking

The protein-small molecule docking can be rigid-rigid molecule docking, rigid protein-flexible ligand docking and flexible side chains in proteins-flexible ligand docking. Untill now, no any methods to simulate flexible backbone of proteins and flexible ligand molecules (induced fitting). After flexible side chains in proteins-flexible ligand docking, it is suggested molecular dynamics calculations. Another problem is the water molecules. In the protein-protein association with decreasing distance between the proteins, in the interface of the two proteins, the structure of the water is changed. Water molecules first help and after hinder the association. No perfect methods are available to predict the conservative water molecule position. The main problem is the knowledge of the binding site in the grid based docking. Docking is constrained to the environment of the supposed binding site which can perform artifacts. In most of the methods, the pharmacophore groups can be assigned for docking, which is a good help to find the best binding molecules.

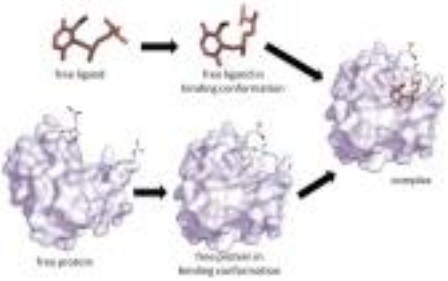


Figure 8.2. A model of protein and ligand in docking

UCSF-DOCK

The protein-ligand docking is based on the (AMBER united atom or all atom) force field scoring. The charges on proteins are AMBER charges or AM1BCC charges (see Chapter 2). The charges on ligands can be AMBER charges, AM1BCC or Gasteiger charges. The Coulomb and van der Waals interactions can be wighted. The score function predicts the best configuration of the associated molecules. One of the first methods was developed as the UCSF-DOCK [7-9]. The shape complementarity is handled by the calculation of solvent accessible surface area (SASA).

On the surface points are generated in equal distance. on the points spheres were placed with the radius of the water molecule (1.4 Å). The spheres are clustered. The user can determine which cluster is the best in docking. Generally, we use the first cluster. In the convex part of the protein (pocket, which can be the binding site) is the best cluster. The shape complementary can be checked by matching, the fitting of heavy (non hydrogen) atoms

to the centre of the spheres (see *Figure 8.3.*). The user can determine the algorithm: manual matching, automated matching, random matching. It is possible, that the largest part of the molecule is fitted first. After fitting this "anchor", the molecule is built up with adding the other parts of the molecule (see *Figure 8.4.*).

For the electrostatic and van der Waals interactions 3D grid files must be generated. The grid spaces are 0.3 Å to 0.5 Å (see *Figure 8.5.*). The ligand molecule can be docked as rigid or flexible molecule. The largest part of the molecule is separated and docked as anchor in the pocket.The protein is rigid. In UCSF-DOCK no flexible side chains and backbones (induced fitting) are available in proteins at all (no flexible backbones are available in docking procedure, the simulation is available by molecular dynamics calculations after docking).

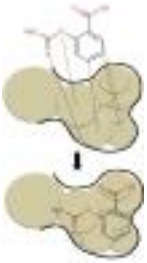


Figure 8.3. The shape complementarity and matching in a pocket (binding site)

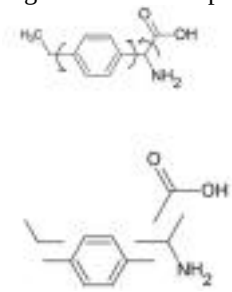


Figure 8.4. The anchor docking in UCSF-DOCK (adapted from the manual of UCSF-DOCK Version 6.5)

Energy scoring is the non-bonding and intramolecular interaction is given by Eq. 8.5.

$$E=\sum_{i=1}^{lig}\sum_{j=1}^{rec}\left(\frac{A_{i,j}}{r_{i,j}^{a}}-\frac{B_{i,j}}{r_{i,j}^{b}}+332\frac{q_i q_j}{Dr_{i,j}}\right) \tag{8.5}$$

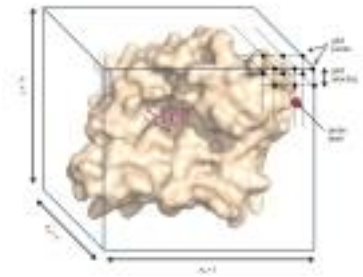


Figure 8.5. The grid based on the electrostatics and van der Waals interactions in the pocket (1yet and a peptide ligand)

Generalization of the van der Waals interactions

$$E_{vdw}=C\varepsilon\left(\frac{2R}{r}\right)^{a}-D\varepsilon\left(\frac{2R}{r}\right)^{b} \tag{8.6}$$

$$E_{vdw} = \varepsilon\left(\frac{b}{a-b}\right)\left(\frac{2R}{r}\right)^{a} - \varepsilon\left(\frac{a}{a-b}\right)\left(\frac{2R}{r}\right)^{b} \tag{8.7}$$

The generalized 12-6 Lennard-Jones equation can be considered as the basic expression between the non-charged atoms (see Eq. 8.4).

$$V_{vdw} = \sum_{i,j}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}}\right) \tag{8.8}$$

r_{ij} is the distance between atoms i and j. *Aij* and *Bij* are parameters derived from the van der Waals parameters of *i* and *j* atoms.

The solvation model was considered by the distance dependent effective dielectric constant and different form of the GB/SA methods (see Chapter 4).

There is a possibility to use penalty function instead of Coulomb and Lennard-Jones 12-6 (or 10-6) function for scoring the association (see *Figure 8.6*.).

During docking, the ligand molecule (anchor) are translated and rotated in the grid box to generate the conformers. The geometry of the complexes are optimized (by e.g. multistep simplex method). Optionally a lot of different conformers (some thousand structures) are generated by translation and rotation of the whole ligand molecules, optimized and ranked on the basis of the value of the score function. There is a possibility to use driver table for a systematic rotation around the torsion in the ligand molecule. The protocol is described in *Figure 8.7*. The algorithm can be used for docking of organic compounds, not only for docking of biomolecules.

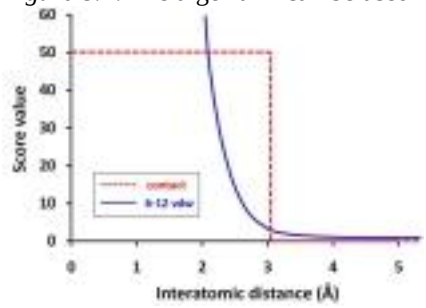


Figure 8.6. Van der Waals interaction (blue) and a penalty function in a contact (red staggered) at 3 Å

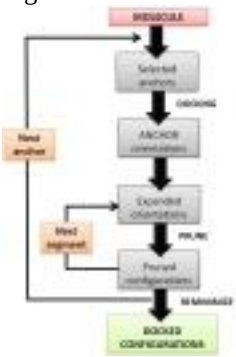


Figure 8.7. The algorithm of docking in UCSF-DOCK (adapted from the manual of UCSF-DOCK Version 6.5)

AUTODOCK

The algorithm is similar to the UCSF-DOCK. It uses AMBER force field with united atom model on protein [10-12]. The charges on ligands were Gasteiger charges. The van der Waals and Coulomb energies were weighted on the basis of experimental binding free energies, so we can obtain the binding free energies of ligand binding. The ligand can be considered rigid or flexible. The protein side chains can be defined as rigid and as flexible partly. The solvation is parametrized for the polar atoms. The protein-ligand complexes are generated by the translation and rotation of the ligand molecule. The score function is calculated by means of electrostatic,

van der Waals and solvation grids. Optimization is possible by the Lamarckian genetic algorithm. After this optimization local optimization is performed. The structures were clustered and ranked of the structures on the basis of the calculated binding free energies. The method is slow and we do not use for virtual screening (see later). The preparation of the input files can be prepared by AUTODOCKTOOLS [12].

An example can be seen about the association of β-sheet breaker peptides a Aβ(1-42) peptide of Alzheimer diseas in Lit. [13].

eHITS

eHITS is (SimBioSys. Ltd., Canada) a very effective method with empirical scoring functions [14]. The different interactions are separated in functions (e.g. π-π stacking between aromatic groups are considered as interactions, see *Figure 1.1*.).

The groups in ligand molecules are separated and independently docked in the binding site. In the end the are connected to each othe on the basis of the original structure (this approach is similar to the fragment based method). The method is suitable for individual ligand docking and it has very good result in virtual screening.

VIRTUAL SCREENING

The aim of the virtual screen (VS) is to find the best scaffold (ligand molecule with shape and electrostatic complementarity) from large databases (100 thousand to 3 million molecules) by rigid docking. One part of the databases are free, one part is commercial the other parts are not available (industrial). On the basis of the score function values, the molecules can be ranked. The best structures can be accepted as the starting point of the drug design. In the docking procedure molecules with high experimental biological effects and with low experimental biological effects are mixed. In the evaluation of the results, we are interested in the enrichment of the molecules with high experimental biological effects and in the end of ranking in the enrichment of the molecules with low experimental biological effects. [15]

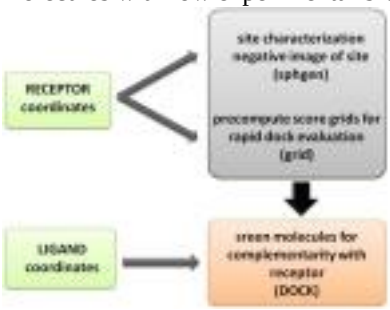


Figure 8.8. The algorithm of docking in UCSF-DOCK (adapted from the manual of UCSF-DOCK Version 6.5)

Some other successful docking methods are available: GLIDE [16] has an excellent score function and the possibility of flexibility of side chains in the proteins with constrained number. FlexX [17] (and its modifications) was developed from the UCSF-DOCK anchor search method with a more reliable score function.

4. Rescoring

After generating the associates in protein-protein and protein-small molecule docking by the score function built in the method, there is a possibility to score again by other, more sophisticated methods. One of the possibility to use force field (CHARMM/ACE, Amber/GBSA, see Chapter 1). Another possibility to use rescoring functions. The basis of Xscore is log P and pKd (tailored score function with different methods together) [18a]. Fred [18b, 18c] includes a lot of score functions (CSScore (Consensus score function with weighted score functions), logP, Gaussian4). After rescoring, the structures are ranked or clustered and ranked.

5. Discovering of Binding Sites

The structure of the proteins are experimentally determined by XRD, synchrorotron technics and NMR. The number of the 3D structures are increasing exponentially (presently ca. 85 thousands structures are available). Sometimes the functions and the binding pockets with different functions are not available. Some methods were developed to find the binding pocket Experimentally, on the basis of NOE results some results were found (see

e.g. [19]). Computational methods were also developed to predict the binding site (e.g. MCSS, GRID [20]). A new method was also developed named CS-MAP [21a]. On the solvent accessible surface equidistant points are generated where organic solvent molecules were placed. The organic solvent molecules were optimized by simplex method by using a simple force field The points can be generated by docking [21b]. The simplex method moves the small molecules on the surface of protein and generate clusters. The binding free energies were calculated by more sophisticated force field (CHARMM/ACE). After clustering the small molecules on the basis of the geometry, the Boltzmann average binding free energy is calculated and ranked. Generally, the first clusters of the small organic solvent molecules show the binding site (consensus binding site) [21d

Motion of the ligand molecules in the binding pocket were studied by 1 ns productive molecular dynamics calculations with GROMACS force fields. The animations with the 10 ps snapshots can be seen in *Figure 8.9.* (methanol), *Figure 8.10.* (acetone), *Figure 8.11.* (urea), *Figure 8.12.* (dimethyl-sulphoxide). The motion of acetone in the binding pocket of thermolysine (2tlx) can *Figure 8.13.*

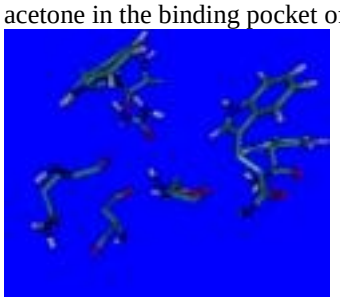


Figure 8.9. Methanol in the binding site of hen egg white lysosime (HEWL) in water

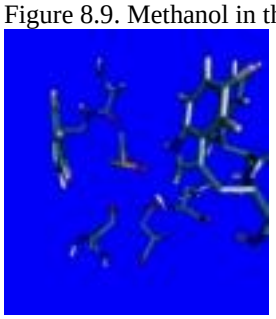


Figure 8.10. Acetone in the binding site of hen egg white lysosime (HEWL) in water

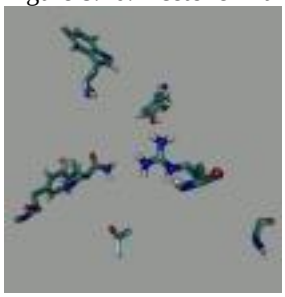


Figure 8.11. Urea in the binding site of hen egg white lysosime (HEWL)

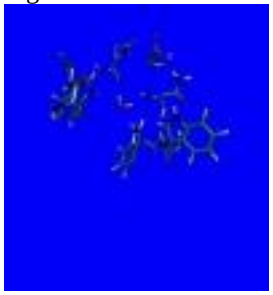


Figure 8.12. Dimethyl-sulphoxide (DMSO) in the binding site of hen egg white lysosime (HEWL)

Figure 8.13. Acetone in the binding site of thermolysine (2tlx)

6. Summary

Docking procedure is the only method to predict computationally the geometries and the binding free energies of the protein/protein and ligand/protein molecules associations. A lot of problems are not cleared yet. We have to accept that the methods have a lot of constraines. The developement of more acceptable methods (considering conservative water molecules, grid calculations, score functions, etc.)

7. References

1. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser, Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. USA, 89, 2195-2199(1972).

2. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser, Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. USA, 89, 2195-2199(1972).

3. http://www.weizmann.ac.il/Chemical_Research_Support/molfit/

4. a) Vakser, I.A., Aflalo, C., Hydrophobic docking: A proposed enhancement to molecular recognition techniques, Proteins, 20, 320-329 (1994). b) Vakser, I.A., Nikiforovich, G.V., Protein docking in the absence of detailed molecular structures, in: Methods in Protein StructureAnalysis (M. Z. Atassi & E. Appella, eds.), Plenum Press, New York, 505-514(1995). c) Vakser, I.A., Protein docking for low-resolution structures, Protein Eng., 8:371- 377(1995..

5. http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/

6. D.W. Ritchie, V. Venkatraman , Ultra-Fast FFT Protein Docking On Graphics Processors. . Bioinformatics, 26, 2398-2405 (2010).. b) G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes, D.W. Ritchie , HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Research, 38, W445-W449 (2010). c) D.W. Ritchie, D. Kozakov, and S. Vajda, Accelerating and Focusing Protein-Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions. Bioinformatics. 24, 1865-1873(2008). d) . http://hex.loria.fr/.

7. a) R. L. DesJarlais and J. S. Dixon, A Shape-and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors. J. Comput-Aided Molec. Design. 8, 231-242, (1994). b) I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, R. and T. E. Ferrin, A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161 269-288 (1982). c) T. J. A. Ewing. and I. D. Kuntz, Critical evaluation of search algorithms for automated molecular docking and database screening. J. Comput. Chem. 18, 1175-1189(1997).

8. a) T.J. A. Ewing, S. Makino, A. G. Skillman, I. D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput-Aided Molec. Design. 15, 411-428 (2001). b) P. T. Lang, S. R. Brozell, S. Mukherjee, E. T. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, I. D. Kuntz, DOCK 6: Combining Techniques to Model RNA-Small Molecule Complexes. RNA 15,1219-1230(2009).

9. http://dock.compbio.ucsf.edu/

10. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Computational Chemistry 2009, 16: 2785-91

11. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998), Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function J. Computational Chemistry, 19: 1639-1662.

12. http://autodock.scripps.edu/

13. C. Hetenyi, Z. Szabo, T. Klement, Z. Datki, T. Kortvelyesi, M. Zarandi, B. Penke, ,Pentapeptide amides interfere with the aggregation of beta-amyloid peptide of Alzheimer's disease, Biochem. and Biophys. Res. Comm. 292, 931-936(2002).

14. a) Zs. Zsoldos, D. Reid, A. Simon, B. S. Sadjad, A. P. Johnson: eHiTS: an innovative approach to the docking and scoring function problems, Current Protein and Peptide Science, 7(5),421-435(2006).b) O. Ravitz, Zs. Zsoldos, A. Simon: Improving molecular docking through eHiTS' tunable scoring function. Journal of Computer-Aided Molecular Design. 2011. c) A, A. Peter Johnson, J. Law, M. Mirzazadeh, O. Ravitz, A. Simon: Computer-aided synthesis design: 40 years on WIREs Comput Mol Sci 2011 001-29 (2011). d) B. Sadjad, Zs. Zsoldos: Toward a Robust Search Method for the Protein-Drug Docking Problem. IEEE/ACM Trans Comput Biol Bioinform. 2010.

15. a) A. Tarcsay, R. Kiss, G. M. Keseru, Site of metabolism prediction on cytochrome P450 2C9: A knowledge-based docking approach, J. Comp.-Aided Mol. Des. 24, 399-408. (2010).

b) G. M Keserű, Lead finding strategies and optimization case studies 2009, Drugs of the Future 35, 143-153. (2010). c) C. G. Ferenczy, G. M.Keserű, Thermodynamics guided lead discovery and optimization, Drug Disc. Today 15, 919-932. (2010). d) J. Huszar, Z. Timar, F. Bogar, B. Penke, R. Kiss, K. K. Szalai, E. Schmidt, A. Papp, G. M. Keseru, Aspartic acid scaffold in bradykinin B1 antagonists, J. Pept. Sci. 15(6), 423-434(2009). d) G. Szabó, R. Kiss, D. Páyer-Lengyel, K. Vukics, J. Szikra, A. Baki, L. Molnár, J. Fischer, G. M. Keserű, Hit-to-lead Optimization of Pyrrolo [1,2-a] quinoxalines as Novel Cannabinoid Type 1 Receptor Antagonists. Bioorg. and Med. Chem. Lett. 19, 3471-3475 (2009). e) R. Kiss, B. Kiss, A. Konczol, F. Szalai, I. Jelinek, V. Laszlo, B. Noszal, A. Falus, G. M. Keseru, Discovery of novel human histamine H4 receptor ligands by large-scale structure-based virtual screening. J. Med. Chem 51(11). 3145-3153(2008).

16. a) T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, J. L. Banks, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening, J. Med. Chem., 47, 1750–1759 (2004). b) R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, D. E. Shaw, M. Shelley, J. K. Perry, P. Francis, P. S. Shenkin, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, J. Med. Chem., 47, 1739–1749 (2004). c) R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, D. T. Mainz, Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes, J. Med. Chem., 49, 6177–6196 (2006). d) N. K. Salam, R. Nuti, W. Sherman, Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis, J. Chem. Inf. Model., 49, 2356–2368 (2009). e) S. Kawatkar, H. Wang, R. Czerminski, D. Joseph-McCarthy, Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide, J. Comput. Aided Mol. Des., 23, 527-539 (2009). f) K. Loving N. K. Salam, W. Sherman, Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation, J. Comput. Aided Mol. Des., 23, 541–554 (2009). g) S. Rao, P. C. Sanschagrin, J. R. Greenwood, M. P. Repasky, W. Sherman, R. Farid, Improving database enrichment through ensemble docking, J. Comput. Aided. Mol. Des., 22, 621-627 (2008). h) K. Loving, I. Alberts, W. Sherman, Computational Approaches for Fragment-Based and De Novo Design, Curr. Top. Med. Chem., 10, 14-32 (2010). i) D.J Osguthorpe,. W. Sherman,A. T. Hagler,, Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering, Chem. Biol. Drug Des., 2012, 80(2), 182-193. j) D. J. Osguthorpe, W. Sherman, A. T. Hagler., Exploring protein flexibility: Incorporating structural ensembles from crystal structures and simulation into virtual screening protocols, J. Phys. Chem. B, (2012). k) M. P. Repasky, R. B. Murphy, J. L. Banks, J. R. Greenwood, I. Tubert-Brohman, S. Bhat, R. A. Friesner, Docking performance of the Glide program as evaluated on the Astex and DUD datasets: A complete set of Glide SP results and selected results for a new scoring function integrating WaterMap and Glide, J. Comput-Aided Mol. Des. 26, 787-799(2012). l) O. Kalid, D. T. Warshaviak, S.

Shechter, W. Sherman, S. Shacham, Consensus Induced Fit Docking (cIFD): methodology, validation, and application to the discovery of novel Crm1 inhibitors, J. Comput-Aided Mol. Des. 26, 1217–1228(20012).

17. M. Rarey M., B. Kramer, T. Lengauer, G. Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol. 261(3):470-89 (1996).

18. a) http://sw16.im.med.umich.edu/software/xtool/manual/usage.html b) M.R. McGann, H.R. Almond, A. Nicholls, J.A. Grant and F.K. Brown, Gaussian docking functions, Biopolymers, 68 (1), 76-90(2003). c) http://www.eyesopen.com/oedocking.

19. E. Liepinsh, G. Otting, Organic solvents identify specific ligand binding sites on protein surfaces. Nature Biotechnology 15, 264-268 (1997).

20. a) A. Caflisch, A. Miranker, M. Karplus, Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase, J. Med. Chem., 36 2142-2167 (1993). b) P. J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem., 28 (7), 849–857(1985).

21. a) S. Dennis, T. Kortvelyesi , S. Vajda, Computational mapping identifies the binding sites of organic solvents on proteins, Proc. of the Natl. Acad. Sci. of the USA. 99(7), 4290-4295(2002). b) T. Kortvelyesi T, M. Silberstein, S. Dennis, S. Vajda, Improved mapping of protein binding sites. J. Comp.-Aided Mol. Design 17(2), 173-186. (2003). c) Improved mapping of protein binding sites, J. Comp.-Aided Mol. Design. 17(2), 173-186 (2003). d) T. Kortvelyesi, S. Dennis, M. Silberstein, L. Brown, S. Vajda, Algorithms for computational solvent mapping of proteins, Proteins-Structure Function and Genetics, Structure and Bioinformatics, 51:(3), 340-351(2003).

8. Further Reading

1. P.T.Lang, D. Moustakas, S. Brozelli, N. Carrascal, S. Mukherjee, T. Balius, S. Pegg, K. Raha, D. Shivakumar, R. Rizzo, D. Case, B. Shhoicet, I. Kuntz, Dock 6.5. Users Manual, University of California, 2006-2012.

2. G. M. Morris, D. S. Goodsell, M. E. Pique, W. L. Lindstrrom, R. Huey, S. Forti, W. E. Hart, S. HallidayR. Belew, A. J. Olson, Autodock Version 4.2, Manual, Autodock4, 2001-2009.

9. Questions

1. Please, describe what is necessary to the protein-protein and the protein-ligand interactions!
2. What is necessary to a real docking procedure?
3. What is the virtual screening? What is the protocol to perform this procedure?
4. What kind of interactions are considered in docking of ligands by UCSF-DOC and AUTODOCK?
5. What is the virtual screening?
6. What is the score function in docking?

10. Glossarry

Docking of proteins to proteins: The protein-to-protein docking fast fourier transformation (FFT) support the shape-to-shape complementarity.

Docking of small molecules to proteins: A lot of structures have to be found to find the best structures which are important in the determination of the acceptable structure.

Chapter 9. Calculation of Ligand-Protein Binding Free Energy

(Tamás Körtvélyesi)

Keywords: ligand-protein binding, binding free energy, classical potential functions, molecular dynamics

1. Introduction

What is described here?This chapter discusses computational methods for estimating the free-energy change accompanied by the binding of small molecules (ligands) to proteins in aqueous solution. The process of binding is analysed in order to rationalize approximate computational schemes. Practical aspect of the methods including computational requirements and accuracy are presented.

What is it used for? Binding free energy is a fundamental quantity in the thermodynamic description of ligand-protein binding and its estimate is widely used at various stages of drug discovery.

What is needed? This chapter assumes the knowledge of basic statistical mechanics and thermodynamics together with the material presented in Chapters 2 (Molecular Mechanics), 4 (Molecular Dynamics) and 8 (Protein-protein and Protein-ligand Binding. Docking methods).

2. Basic Equations of Binding Thermodynamics

The binding affinity of a ligand (L) to a protein (P) can be characterized by the dissociation constant, K_d

$$K_d = \frac{[L][P]}{[LP]} \tag{9.1}$$

corresponding to the process

$$LP = L + P \tag{9.2}$$

The logarithm of the dissociation constant is proportional to the Gibbs free energy of binding (ΔG_{bind}).

$$\Delta G_{bind} = RT \ln K_d \tag{9.3}$$

where R is the universal gas constant and T is the absolute temperature. ΔG_{bind} is a function of the binding enthalpy (ΔH_{bind}) and the binding entropy (ΔS_{bind}).

$$\Delta G_{bind} = \Delta H_{bind} - T\Delta S_{bind} \tag{9.4}$$

The above equations show that an improved binding affinity, i.e. a decreased Kd is equivalent with a decreased – more negative – binding free energy. This can be achieved with a more negative enthalpy and with increased entropy change.

3. Decomposition of the Binding Process. The role of solvent

Ligand binding signifies the process in which the originally separated and solvated ligand and protein form the solvated ligand-protein complex. This process can be decomposed – at least in theory - into several steps. A usual decomposition includes desolvation of the ligand and the binding site, changing the conformation of both the ligand and the protein and forming interactions between them. Desolvation restructures water around the ligand that results in a significant entropic reward. Replacement of water from the binding site may have different enthalpic and entropic consequences depending on the binding interactions of the replaced water molecules. Binding is usually accompanied by conformational rearrangement of both the ligand and the receptor and this represents an enthalpic penalty in most cases. Formation of the ligand-receptor complex is typically

coupled to forming new interactions between the ligand and its binding site that are enthalpically beneficial. Molecular recognition of the ligand, however, limits its external rotational and translational freedom as well as ligand and protein flexibility and therefore represents an entropic penalty. Although the thermodynamic impact of long range effects is usually neglected they could also contribute to ligand binding.

It is observed for a great variety of systems that structural variations resulting in small changes in ΔG imply more significant changes in its components, ΔH and TΔS. This phenomenon is referred to as enthalpy-entropy compensation. It is manifested also in the wider spread of observed ΔH and TΔS than ΔG values as it is illustrated in *Figure 9.1.*

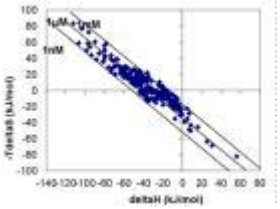


Figure 9.1. Binding enthalpy and entropy values for 285 ligand-protein complexes. Adapted from G.G. Ferenczy , G. M. Keserű J. Chem. Inf. Model. 2012, 52, 1039.

An interpretation of the compensation hypothesizes that an enthalpically more favorable binding imposes a more severe restriction to the motion of the interacting partners and thus a more significant unfavorable entropic change (see e.g. ref. [1]). Water models that are able to account for the enthalpy-entropy compensation of aqueous processes have also been proposed [2] . It is worth mentioning that although the thermodynamics of host-guest complexes are different in water and in organic solvents the enthalpy-entropy compensation is not restricted to aqueous systems [3] .

A detailed understanding of water properties, hydration and hydrophobicity is essential to the rationalization of binding thermodynamics in biological systems . Water is a highly complex liquid. No comprehensive theory is available to explain all experimental observations and to adequately describe aqueous processes at the level of molecular detail. A particularly intriguing phenomenon is the hydrophobic effect that refers to the transfer of apolar compounds, either from their liquid state or from a solution in an apolar solvent, to water [4]. This is a process that includes the disruption of interactions between the apolar compound and its apolar environment, refilling the vacancy in the apolar medium, cavity formation in liquid water, the insertion of the nonpolar solute, the onset of the solute-solvent interactions and the reordering of the water molecules in the close proximity of the solute. This process is accompanied by a free energy increase. At room temperature, both the enthalpy and the entropy are negative and the later dominates. Increasing the temperature the free energy hardly changes but the high and positive heat capacity of hydration implies that the entropy driven process at room temperature becomes enthalpy driven at higher temperature. Theories of the hydrophobic effect at the level of molecular structure usually concentrate on those steps that involve water (i.e. cavity formation in water, placement of the solute into the cavity and the restructuring of the water around the solute) and are termed hydrophobic hydration. These theories have been elaborated basically along two lines. One argues that the small size of water molecules is a key feature in producing negative entropy in opening up a cavity for the solute molecule[5],[6],[7]. The other is based on the hydrogen bonding properties of water and assumes different hydrogen bonding in the bulk than in contact with a solute (mixture or two-state water models; see ref. 11 and references therein). A combination of these factors was also proposed as the origin of hydrophobic effect[8],[9],[10],[11].

Protein-ligand complex formation in water is a related but more complex process than hydrophobic hydration; the latter can serve as a model for certain aspects of the former. Ligand binding shows resemblance to micelle formation in the sense that both processes are associated with the coalescence of solutes and thus a decrease of cavity size and a release of solvating water molecules [12],[13],[14]. Thus ligand binding is typically accompanied by desolvation of hydrophobic groups with the corresponding thermodynamic signatures. These include entropy increase and a negative heat capacity at constant pressure, the contributing factors to the latter being debated [15].

Ligand binding to protein is usually accompanied by conformational changes of both partners. These changes are associated with unfavorable free-energy that is counterbalanced by favorable contributions of the binding. Structural changes of proteins upon ligand binding is often identified e.g. by the comparison of the X-ray structures of the ligand free (apo) protein with that found in the complex. However, the observed crystal structures give no direct information for the free-energy change of the complex formation for several reasons. The protein is only a part of the whole system and the examination of the free-energy consequence of its conformational change has limited significance. Moreover, X-ray produces crystal packing biased snapshots of

dynamic structures and these – in the best case - may give estimates of the enthalpy, but not of the free-energy change. Similarly, the ligand often binds in a conformation with higher energy (enthalpy) than that of the minimum energy solvated molecule.

The ligand and the protein form new interactions upon complex formation. It is important to realize that when the ligand binds to the protein then ligand-water and protein-water contacts are replaced by ligand-protein and water-water contacts the latter are formed by some water molecules participated in the solvation before, and become part of the bulk solvent after the binding. In this way, the newly formed ligand-protein and water-water interactions replace those existed before the binding. Thus a net free energy gain can only be achieved when good steric and electrostatic complementarity between the ligand and the protein is realized.

4. Molecular Dynamics Based Computational Methods

The Helmholtz free energy (F) expressed as

$$F = -kT \ln Z \tag{9.5}$$

where the partition function

$$Z = \iint \exp\left(\frac{-E}{kT}\right) dp\, dr \tag{9.6}$$

includes an integral over the states of the system. Here k is the Boltzmann constant, T is the absolute temperature and E is the energy corresponding to state with position vector r and momentum vector p. The evaluation of the partition function for systems as complex as those in ligand-protein binding is not feasible. Even its approximation with a set of Boltzmann weighted snapshots - generated from Monte Carlo (MC) or molecular dynamics (MD) simulations - is highly inaccurate and therefore various methods have been worked out to estimate free energies or free energy differences without the evaluation of Z.

Note, that Gibbs free energy changes (ΔG) in solution are assumed to be well approximated by Helmholtz free energy changes (ΔF) and this is exploited in the following discussion.

Binding free energy differences can be obtained with a reasonable amount of computational work by alchemical transformations. An example is presented in *Figure 9.2*.

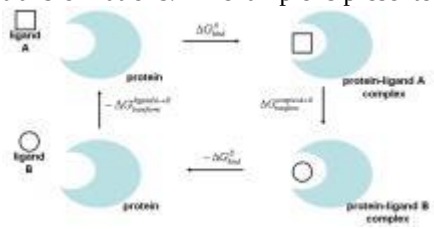


Figure 9.2. Thermodynamic cycle for calculating the difference in binding free-energies of two ligands (taken from G.G. Ferenczy and G. M. Keserű in Physico-Chemical and Computational Approaches to Drug Discovery, J. Luque, X. Barril Eds., The Royal Society of Chemistry, Cambridge 2012, pp 23-79. Reproduced by permission of The Royal Society of Chemistry

The free energy change in this thermodynamic cycle is zero. Thus the difference of the binding free energies of ligands A and B can be written as

$$\Delta\Delta G = \Delta G^{A}_{bind} - \Delta G^{B}_{bind} = \Delta G^{ligand\ A \to B}_{transform} - \Delta G^{complex\ A \to B}_{transform} \tag{9.7}$$

This equation shows that the binding free energy difference of ligands can be calculated as the difference of the free energies of two alchemical transformations; one that transforms the unbound solvated ligand A into B, and another that transforms the solvated protein-ligand A complex into protein-ligand B complex. The advantage of treating these alchemical transformations is that they connect systems whose free energy difference can be calculated with improved efficiency.

The two most widely used methods for calculating free energy differences of alchemical transformations are thermodynamic integration (TI) and free-energy perturbation (FEP). The TI equation has a particularly simple form when the potential functions of the two states are linear in a parameter λ. This is illustrated in *Figure 9.3*.

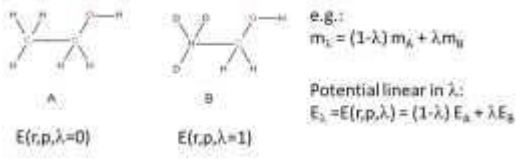


Figure 9.3. Transformation of ethanol into methanol (D represents dummy atoms) assuming a potential linear in the transformation parameter λ

The free energy difference corresponding to the transformation in *Figure 9.3.* can be written as

$$\Delta F = F_B - F_A = F(\lambda=1) - F(\lambda=0) = \int_{\lambda=0}^{\lambda=1} F'(\lambda)\,d\lambda = \frac{d}{d\lambda}[-kT \ln Z(\lambda)] = \left\langle \frac{\partial E}{\partial \lambda} \right\rangle \tag{9.8}$$

where the rightmost formula obtained by invoking Eq. (9.6) and contains the ensemble average of $\langle \partial E/\partial \lambda \rangle$ over the λ distribution of states. Owing to the linear dependence of the energy on the parameter λ, the free energy difference simplifies to

$$\Delta F = F_B - F_A = \int_0^1 \langle E_B - E_A \rangle_\lambda \, d\lambda \tag{9.9}$$

TI calculations include multiple simulations with different λ values and a numerical integration over λ to obtain the free energy difference.

The basic equation of FEP[16] for the free energy difference of two states can be obtained in the following way. The free energy difference is written as

$$F_B - F_A = -kT \ln \left[\frac{\iint \exp\left(-\frac{E_B}{kT}\right) dp\,dr}{\iint \exp\left(-\frac{E_A}{kT}\right) dp\,dr} \right] \tag{9.10}$$

Inserting $1=\exp(-E_A/kT)\exp(E_A/kT)$ in the numerator

$$F_B - F_A = -kT \ln \left[\frac{\iint \exp\left(-\frac{E_A}{kT}\right) \exp\left(\frac{E_A}{kT}\right) \exp\left(-\frac{E_B}{kT}\right) dp\,dr}{\iint \exp\left(-\frac{E_A}{kT}\right) dp\,dr} \right] \tag{9.11}$$

and replacing E_B-E_A by ΔE

$$F_B - F_A = -kT \ln \left[\frac{\iint \exp\left(-\frac{E_A}{kT}\right) \exp\left(-\frac{\Delta E}{kT}\right) dp\,dr}{\iint \exp\left(-\frac{E_A}{kT}\right) dp\,dr} \right] \tag{9.12}$$

we obtain the ensemble average of exp(-ΔE/kT) over the initial state. This can be written as

$$\Delta F = F_B - F_A = -kT \ln \left\langle \exp\left(\frac{E_B - E_A}{kT}\right) \right\rangle_A \tag{9.13}$$

where the $\langle\rangle_A$ brackets indicate ensemble average over system A. A FEP calculation includes the evaluation of the energy difference between the two states and the ensemble average is taken over the first state. Improved accuracy via a better sampling can be achieved by dividing the transition between the two end states to several steps. Performing also a backward simulation allows an estimation of the convergence.

Non-equilibrium work methods[17],[18] are related approaches based on the equality of the work associated to the non-equilibrium switch between two states and the free energy difference of these state

$$\Delta F = F_2 - F_1 = -kT \ln \left\langle \exp\left(-\frac{W}{kT}\right) \right\rangle \tag{9.14}$$

where W is the external work performed on the system and the average is taken along the possible trajectories.

The double-decoupling method[19] deserves special attention as it is able to calculate standard binding free energies. The thermodynamic basis of the methods is shown in *Figure 9.3*.

Figure 9.4. Thermodynamic basis of the double-decoupling method. (Taken with permission from ref. [19].)

Double decoupling includes two simulations; one with the ligand in solution and another with the ligand together with the protein in solution. In both simulations the interactions of the ligand with its environment is decoupled.

We do not discuss here the various technical aspects of free energy simulations but we note that sophisticated techniques are required to improve sampling and data analysis and to obtain meaningful estimations of binding free energies or their differences. Interested readers are referred to recent reviews.[20],[21],[22]

The quality of the potential energy function applied in the simulation is a crucial determinant of the accuracy of the free energy estimation. Most calculations use classical force fields. These give a reasonable description of the proteins owing to the limited variability of protein sequences but they may be less appropriate for diverse ligands or cofactors. A particularly challenging aspect of force fields is the proper account of polar interactions. The evaluation of long range electrostatic interactions is computationally demanding and their best approximations invoke either periodicity or a dielectric continuum beyond a certain cutoff distance. A proper description of polar interactions is problematic also at short interatomic separations. The oversimplified representation of molecular charge densities by atomic point charges and the neglect of polarization may affect the quality of the description adversely.

All methods described above aim at estimating the binding free energy. In case, we wish to calculate its enthalpy and entropy components we are facing with additional difficulties. In principal, the enthalpy should be easier to evaluate than the free energy owing to the smaller fluctuations of the ensemble averages of the former (see e.g. ref. 70). However, these fluctuations are still too high to obtain meaningful results with reasonable computational effort. For the same reason, the evaluation of an enthalpy difference as the difference of ensemble averages is highly inaccurate. Various methods have been proposed to calculate enthalpy or entropy differences as ensemble averages (rather than the difference of ensemble averages). These methods are typically based on formulas for free energy differences and exploit relationships between thermodynamic quantities.[23],[24] Unfortunately, they are unable to achieve the accuracy of or best techniques for evaluating free energy differences (vide supra); they give results with reasonable accuracy for simple solute-solvent systems but they are not appropriate for treating ligand-protein binding.

5. Other Computational Methods

5.1. Estimation of the Free Energy

Molecular simulation based methods give a theoretically well founded and potentially accurate description of ligand binding thermodynamics. On the other hand, primarily for practical reasons partly discussed above, they do not offer a general solution to calculate binding free energies and their components. This prompted the development of a plethora of other methods to calculate the binding free energy or its specific contributions. Some of them include simulation based estimate of certain properties but they invoke additional approximations with respect to the methods discussed in section 9.5.

MM-PBSA[25] calculates the binding free energy as the difference between the free energy of the solvated complex and those of the solvated unbound components. The free energy is approximated with the following terms

$$G = E_{MM} + G_{PBSA} - TS_{MM} \tag{9.15}$$

E_{MM} is the molecular mechanical energy, G_{PBSA} is the solvation free energy and TS_{MM} is the solute entropy. Several variants for the calculation of these terms have been proposed.

E_{MM} can come from simple molecular mechanical minimization or from molecular dynamics trajectories. In this latter case the energy of the unbound molecules can be obtained from simulations performed for the unbound molecules or from the simulation performed for the complex. G_{PBSA} is calculated with a numerical solution of the Poisson-Boltzmann equation and an estimate of the nonpolar free energy with a surface area term. TS_{MM} usually includes an estimate of the conformational entropy obtained by normal-mode analysis. MM-PBSA was found to be appropriate to improve virtual screening results when applied as a post-docking filter and also to prioritize design compounds. On the other hand, it is expected to correctly rank compounds with free energy differences of at least 3 kcal/mol, at best.[26]

The Linear Interaction Energy (LIE) method[27],[28] estimates the standard binding free energy as the sum of an electrostatic (E^{el}) and a van der Waals (E^{vdw}) term

$$\Delta G=\frac{1}{2}\left(\langle E^{el}\rangle_{bound}-\langle E^{el}\rangle_{free}\right)+\alpha\left(\langle E^{vdw}\rangle_{bound}-\langle E^{vdw}\rangle_{free}\right) \qquad (9.16)$$

Where ½ comes from the assumption of linear response and a is an adjustable parameter. Again, several variants of the method have been proposed. They include the replacement of the ½ coefficient of the electrostatic term by an adjustable parameter[29], the addition of a term proportional with the solvent accessible surface area to account for cavity formation[30],[31] and the replacement of the molecular mechanical electrostatic and van der Waals energy terms by quantum mechanical/molecular mechanical (QM/MM) interaction energy calculated for the time averaged structure.[32] The application of the LIE method requires the knowledge of some binding free energies to perform the calibration of the adjustable parameters. These linear response based methods were shown to give reasonable results for certain series of ligands. In other cases, LIE estimations are subject to important errors and owing to the approximations involved an a priori assessment of the quality of the results is difficult if at all possible.

Scoring functions, designed for a fast ranking of ligand-protein complexes also estimate binding free energies (see Chapter 8 and refs.[33] and [34] for recent reviews). In typical applications scores for a large number of ligands complexed with the same protein are calculated and then a selection of top ranked ligands gives a set enriched with compounds showing reasonable binding affinity towards the protein. Various schemes are used to derive scoring functions and they largely differ in the way the various free energy components are approximated. As scoring functions are typically used for treating a large number of compounds (often in the range of 10^5-10^6) accuracy and rigour in the theoretical foundation are sacrificed for speed. As a consequence, the correlation between scores and binding free energies is poor, and the enthalpy and entropy components cannot be straightforwardly identified. On the other hand, with the improvement of methodology and with the advancement of available computer power the notion of scoring function starts to expand and to include more involved methods.

5.2. Estimation of the Enthalpy

Quantum mechanics (QM) offers a potentially highly accurate description of intermolecular interactions. Its advantages over molecular mechanics (MM) include that no parameterization is required and thus compounds with unusual structural motifs can be treated, and the accuracy of the description can be systematically increased. Unfortunately, the high computational demand represents a serious limitation to sampling the configurational space by QM. At the same time, a limitation of the QM evaluation of the energy of configurations generated by MM stems from the differences of the QM and MM potential surfaces[35]. An alternative approach is the approximation of the enthalpy of binding by semiempirical QM calculations for a single configuration. A final remark concerning the application of QM methods is that highly accurate description of selected factors does not necessarily results in higher quality thermodynamic quantities that emerge as the sum of several partially cancelling contributions.[36]

The thermodynamic characterization of ligand protein binding using structural data with an empirical parameterization stems from the successful application of this type of approach for the description of protein folding (see e.g. ref. [37]). Key parameters in predicting thermodynamic quantities, ΔG, ΔH, ΔS and ΔC_p, in protein folding are the change of solvent accessible surface areas (ΔASA) and its dissection into apolar (ΔASA_{apolar}) and polar parts (ΔASA_{polar}). These descriptors with parameterization specific for ligand-protein binding enthalpy were used in applications to ligand protein binding.[38],[39] An important simplification of this approach is that it does not include an explicit term for the ligand-protein interaction rather the contribution of

this interaction to the enthalpy is implicit in the surface area terms. This approximation is unlikely to be able to reflect the known sensitivity of the enthalpy to the geometry of the interacting partners. [40]

5.3. Estimation of the Entropy

Various approximate methods have been proposed to estimate the entropy change upon ligand binding. They typically address certain components of the entropy, most often the configurational entropy change of the solute. These methods cannot be directly compared to experimental results as they do not provide us with measurable quantities. It is also important to realize, that the usual decomposition of the entropy into translational, rotational and vibrational components as $S=S_{trans}+S_{rot}+S_{vib}$is somewhat arbitrary. Similarly, the hard (bond length and angles) and soft (dihedral and external) coordinates may couple and it is an approximation to treat selected components separately.

Normal mode analysis estimates the entropy from an energy minimized structure assuming harmonic potentials.[41],[42],[43] The value of the calculated $T\Delta S$ was found to vary on the selected minimized structure by 5kcal/mol in unfavorable cases.[44] Another factor affecting the utility of the normal mode analysis is the validity of the harmonic approximation.

The quasiharmonic (QH) method calculates the configurational entropy assuming a multivariate Gaussian distribution for the Boltzmann probabilities and deriving the covariance matrix of the coordinates from computer simulations.[45],[46] Shortcomings of the QH method as applied to biochemical systems include an overestimation of the entropy and slow convergence.[47],[48],[49]

The "mining minima" approach[50], [51], [52] is able to estimate configurational entropy and is exempt from assumptions of the previous methods. It identifies local minima of the potential surface, i.e. predominant low-energy conformations and their contributions to the configurational integral is evaluated taking anharmonicity also into account. The computational intensive search is currently practical with implicit solvent models only.

6. References

1. D.H. Williams, E. Stephens, D.P. O'Brien, M. Zhou, „Understanding Noncovalent Interactions: Ligand Binding Energy and Catalytic Efficiency from Ligand-Induced Reductions in Motion within Receptors and Enzymes", Angew. Chem. Int. Ed., 43, 6596–6616, (2004).

2. B. Lee, G. Graziano, „A Two-State Model of Hydrophobic Hydration That Produces Compensating Enthalpy and Entropy Changes", J. Am. Chem. Soc., 118, 5163-5168, (1996).

3. K.N. Houk, A.G. Leach, S.P. Kim, X. Zhang, „Binding Affinities of Host–Guest, Protein–Ligand, and Protein–Transition-State Complexes", Angew. Chem. Int. Ed., 42, 4872-4897 (2003).

4. W. Blokzijl J.B.F.N. Engberts, „Hydrophobic Effects. Opinions and Facts", Angew. Chem. Int. Ed., 32, 1545–1579, (1993).

5. M. Lucas, „Size effect in transfer of nonpolar solutes from gas or solvent to another solvent with a view on hydrophobic behavior", J. Phys. Chem., 80, 359-362, (1976).

6. B. Lee, „The physical origin of the low solubility of nonpolar solutes in water", Biopolymers, 24, 813-823, (1985).

7. B. Lee, „Enthalpy-entropy compensation in the thermodynamics of hydrophobicity", Biophys. Chem., 51, 271-278, (1994).

8. M. Kinoshita, "Molecular origin of the hydrophobic effect: Analysis using the angle-dependent integral equation theory", J. Chem. Phys., 128, 024507 (2008).

9. T. Lazaridis, „Solvent Reorganization Energy and Entropy in Hydrophobic Hydration", J. Phys. Chem. B, 104, 4964-4979, (2000).

10. T. Lazaridis, "Solvent Size vs Cohesive Energy as the Origin of Hydrophobicity", Acc. Chem. Res., 34, 931-937, (2001).

11. N. Muller, “Search for a realistic view of hydrophobic effects”, Acc. Chem. Res., 23, 23-28, (1990).

12. D. Chandler, “Interfaces and the driving force of hydrophobic assembly”, Nature, 437, 640-647, (2005).

13. E. Fisicaro, C. Compari, A. Braibanti, “Entropy/enthalpy compensation: hydrophobic effect, micelles and protein complexes”, Phys. Chem. Chem. Phys., 6, 4156-4166, (2004).

14. E. Fisicaro, C. Compari, A. Braibanti, „Hydrophobic hydration processes: General thermodynamic model by thermal equivalent dilution determinations”, Biophys. Chem., 151, 119-138, (2010).

15. A. Cooper, „Heat capacity effects in protein folding and ligand binding: a re-evaluation of the role of water in biomolecular thermodynamics”, Biophys Chem., 115, 89-97, (2005).

16. R. W. Zwanzig, “High- Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases”, J. Chem. Phys. 22, 1420-1426, (1954).

17. C. Jarzynski, “Nonequilibrium Equality for Free Energy Differences”, Phys. Rev. Lett., 78, 2690-2693, (1997).

18. C. Jarzynski, “Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach”, Phys. Rev. E., 56, 5018-5035, (1997).

19. M. K. Gilson, J. A. Given, B. L. Bush, J. A. McCammon, „The statistical-thermodynamic basis for computation of binding affinities: a critical review”, Biophys. J., 72, 1047-1069, (1997).

20. A. Pohorille, C. Jarzynski, C. Chipot, „Good Practices in Free-Energy Calculations”, J. Phys. Chem. B, 114, 10235-10253 (2010).

21. T. Steinbrecher, A. Labahn, “Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies”, Curr. Med. Chem., 17, 767-785, (2010).

22. J. Michel, J. W. Essex, “Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations”, J. Comput. Aided Mol. Des., 24, 639-658, (2010).

23. N. Lu, D.A. Kofke, T.B. Woolf, „Staging Is More Important than Perturbation Method for Computation of Enthalpy and Entropy Changes in Complex Systems”, J. Phys. Chem. B, 107, 5598-5611, (2003).

24. C. Peter, C. Oostenbrink, A. van Dorp, W.F. van Gunsteren, “Estimating entropies from molecular dynamics simulations” J. Chem. Phys., 120, 2652-2661, (2004).

25. P.A. Kollman, I. Massova, C. Reyes; B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham III, “Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models”, Acc. Chem. Res., 33, 889-897, (2000).

26. B. Kuhn, P. Gerber, T. Schulz-Gasch, M. Stahl, “Validation and Use of the MM-PBSA Approach for Drug Discovery”, J. Med. Chem., 48, 4040-4048, (2005).

27. J. Åqvist, C. Medina, J.–E. Samuelsson, “A new method for predicting binding affinity in computer-aided drug design”, Protein Eng., 7, 385-391, (1994).

28. J. Åqvist, V.B. Luzhkov, B.O. Brandsdal, “Ligand Binding Affinities from MD Simulations”, Acc. Chem. Res., 35, 358-365, (2002).

29. T. Hansson, J. Marelius, J. Åqvist, “Ligand binding affinity prediction by linear interaction energy methods”, J. Comput. Aided Mol. Des., 12, 27-35, (1998).

30. H.A. Carlson, W.L. Jorgensen, „An Extended Linear Response Method for Determining Free Energies of Hydration”, J. Phys. Chem., 99, 10667-10673, (1995).

31. D.K. Jones-Hertzog, W.L. Jorgensen, "Binding Affinities for Sulfonamide Inhibitors with Human Thrombin Using Monte Carlo Simulations with a Linear Response Method", J. Med. Chem., 40, 1539-1549, (1997).

32. A. Khandelwal, V. Lukacova, D. Comez, D.M. Kroll, S. Raha, S. Balaz, "A Combination of Docking, QM/MM Methods, and MD Simulation for Binding Affinity Estimation of Metalloprotein Ligands", J. Med. Chem., 48, 5437-5447, (2005).

33. R. Rajamani, A. C. Good, "Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development.", Curr. Opin. Drug Discov. Devel., 10, 308-315, (2007).

34. T. Cheng, X. Li, Y, Li, Z. Liu, R. Wang, "Comparative Assessment of Scoring Functions on a Diverse Test Set", J. Chem. Inf. Model., 49, 1079-1093, (2009).

35. A. Weis, K. Katebzadeh, P. Söderhjelm, I. Nilsson, U. Ryde, "Ligand Affinities Predicted with the MM/PBSA Method: Dependence on the Simulation Method and the Force Field", J. Med. Chem., 49, 6596-6606, (2006).

36. P. Söderhjelm, J. Kongsted, S. Genheden, U. Ryde, "Estimates of ligand-binding affinities supported by quantum mechanical methods", Interdisp. Sci, Compout. Life Sci., 2, 21-37, (2010).

37. K.P. Murphy, V. Bhakuni, D. Xie, E. Freire, "Molecular basis of co-operativity in protein folding: III. Structural identification of cooperative folding units and folding intermediates", J. Mol. Biol., 227, 293-306, (1992).

38. B.M. Baker, K.M. Murphy, "Prediction of binding energetics from structure using empirical parameterization", Methods Enzym., 295, 294-315 (1998).

39. I. Luque, E. Freire, "Structural parameterization of the binding enthalpy of small ligands", Proteins, 49, 181-190, (2002).

40. E. Freire, "Do enthalpy and entropy distinguish first in class from best in class?", Drug Discov. Today, 13, 869-874, (2008).

41. N. Gō, H. A. Scheraga, "Analysis of the Contribution of Internal Vibrations to the Statistical Weights of Equilibrium Conformations of Macromolecules", J. Chem. Phys., 51, 4751-4766, (1969).

42. N. Gō, H. A. Scheraga, "On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation", Macromolecules, 9, 535-542, (1976).

43. D. A. Case, "Normal mode analysis of protein dynamics", Curr. Opin, Struct. Biol., 4, 285-290, (1994).

44. B. Kuhn, P. A. Kollman, "Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models", J. Med. Chem., 43, 3786-3791, (2000).

45. M. Karplus, J. N. Kusshick, "Method for estimating the configurational entropy of macromolecules", Macromolecules, 14, 325-332, (1981).

46. M. M. Teeter, D.A. Case, "Harmonic and quasiharmonic descriptions of crambin ", J. Phys. Chem, 94, 8091-8097, (1990).

47. C-E. Chang, W. Chen, M. K. Gilson, "Evaluating the Accuracy of the Quasiharmonic Approximation ", J. Chem. Theory Comput., 1, 1017-1028, (2005).

48. H. Gohlke, D. A. Case, "Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf", J. Comput. Chem., 25, 238-250, (2004).

49. S-T. D. Hsu, C. Peter, W. F. van Gunsteren, A. Bonvin, „Entropy Calculation of HIV-1 Env gp120, its Receptor CD4, and their Complex: An Analysis of Configurational Entropy Changes upon Complexation", Biophys J., 88, 15-24, (2005).

50. M.S. Head, J. A. Given, M. K. Gilson, „"Mining Minima": Direct Computation of Conformational Free Energy", J. Phys. Chem. A, 101, 1609-1618, (1997).

51. C-E. Chang, M. K. Gilson, „Free Energy, Entropy, and Induced Fit in Host−Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm" J. Am. Chem. Soc., 126, 13156-13164, (2004).

52. C-E. Chang, W. Chen, M. K. Gilson, "Ligand configurational entropy and protein binding", Proc. Natl. Acad. Sci. USA, 104, 1534-1539, (2007).

7. Further Readings

1. Chipot, C.; Pohorille, A.(eds.) Free Energy Calculations: Theory and Applications in Chemistry and Biology. Springer Series in Chemical Physics 86. Springer, Berlin Heidelberg 2007.

2. Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J. Comput. Aided Mol. Des. 24, 639-658, (2010).

3. Chodera, J. D.; Mobley, D. L., Shirts, M. L. Dixon, R. W., Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. Curr. Opin. Struct. Biol., 21, 1-11, (2011).

8. Questions

1. What is the relationship between the ligand-protein dissociation constant (Kd) and the binding free energy (ΔG)?

2. Why is it advantageous to calculate the difference of binding free energies of two ligands rather than the binding free energies of the individual ligands.

3. What is the sign of the contribution of conformational change to the enthalpy of binding?

4. How does water rearrangement contribute to the entropy change upon ligand-protein binding?

5. What does alchemical transformation mean?

Chapter 10. Introduction to Cheminformatics. Databases.

(Róbert Rajkó, Tamás Körtvélyesi)

Keywords:

What is described here? Chem(o)informatics is defined, and some basic and advanced statistical methods used for cheminformatics are described.

What is it used for? Building Structure–Activity Relationships/Quantitative Structure–Activity Relationships/Quantitative Structure–Property Relationships (SAR/QSAR/QSPR) for modeling bioactivity based on special descriptors.

What is needed? Some basic theoretical chemistry knowledge, uni- and multivariate statistics, matrix algebra.

1. Introduction

Chem(o)informatics is for developing models linking chemical structure and various molecular properties. In this sense cheminformatics relates to two other modeling approaches – quantum chemistry and force-field simulations. These three complementary fields differ with respect to the form of their molecular models, their basic concepts, inference mechanisms and domains of application. Unlike the molecular models used in quantum mechanics (ensembles of nuclei and electrons) and force field molecular modeling (ensembles of "classical" atoms and bonds), cheminformatics treats molecules as molecular graphs or related descriptor vectors with associated features (physicochemical properties, biological activity, 3D geometry, etc.) [Varnek2011]. The ensemble of graphs or descriptor vectors forms a chemical space in which some relations between the objects must be defined. Unlike real physical space, a chemical space is not unique: each ensemble of graphs and descriptors defines its own chemical space. Thus, cheminformatics could be defined as a scientific field based on the representation of molecules as objects (graphs or vectors) in a chemical space [Varnek2011].

Cheminformatics considers a molecule as a graph or an ensemble of descriptors generated from this graph. A set of molecules forms a chemical space for which the relationships between the objects themselves, on one hand, and between their chemical structures and related properties, on the other hand, are established using two main mathematical approaches: graph theory and statistical learning. Due to the rapidity of such calculations, these structure-property relationships can be applied to fast screening of large databases [Varnek2011].

2. Basic Statistical Methods

Among a multitude of descriptors currently used in Structure–Activity Relationships/Quantitative Structure–Activity Relationships/Quantitative Structure–Property Relationships (SAR/QSAR/QSPR) studies, fragment descriptors (application as atoms and bonds increments in the framework of additive schemes) occupy a special place [Baskin2008].

The epoch of QSAR (Quantitative Structure–Activity Relationships) studies began in 1963–1964 with two seminal approaches: the σ-ρ-π analysis of Hansch and Fujita [Hansch1963, Hansch1964] and the Free–Wilson method [Free1964]. The former approach involves three types of descriptors related to electronic, steric and hydrophobic characteristics of substituents, whereas the latter considers the substituents themselves as descriptors. Both approaches are confined to strictly congeneric series of compounds. The Free–Wilson method additionally requires all types of substituents to be sufficiently present in the training set. A combination of these two approaches has led to QSAR models involving indicator variables, which indicate the presence of some structural fragments in molecules.

In organic chemistry, decomposition of molecules into substituents and molecular frameworks is a natural way to characterize molecular structures. In QSAR, both the Hansch–Fujita [Hansch1963, Hansch1964] and the Free–Wilson [Free1964] classical approaches are based on this decomposition, but only the second one explicitly accounts for the presence or the absence of substituent(s) attached to molecular framework at a certain position. While the multiple linear regression technique was associated with the Free–Wilson method, recent

modifications of this approach involve more sophisticated statistical and machine-learning approaches, such as the principal component analysis [Fleischer2000] and neural networks [Hatrik1996]. Disconnected atoms represent the simplest type of fragments. Usually, the atom types account for not only the type of chemical element but also hybridization, the number of attached hydrogen atoms (for heavy elements), occurrence in some groups or aromatic systems, etc. Nowadays, atom-based methods are used to predict some physicochemical properties and biological activities. Chemical bonds are another type of simple fragment. Topological torsions are defined as a linear sequence of four consecutively bonded non-hydrogen atoms. The above-mentioned structural fragments – atoms, bonds and topological torsions – can be regarded as chains of different lengths.

They are used to assess a chemical or biological property *P* in the framework of an additive scheme based on chainlike contributions:

$$P=\sum_{i=1}^{N} n_i C_i \qquad (10.1)$$

where n_i is the number of atoms, bonds or topological torsions of *i*-type, C_i is corresponding chainlike contributions.

Hansch pioneered the use of descriptors related to a molecule's electronic characteristics and to its hydrophobicity [Leach2007]. This led him to propose that biological activity could be related to the molecular structure via equations of the following form:

:

$$\log\left(\frac{1}{C}\right)=k_1 \log P+k_2\sigma+k_3 \qquad (10.2)$$

where *C* is the concentration of compound required to produce a standard response in a given time, log *P* is the logarithm of the molecule's partition coefficient between 1-octanol and water and σ is the appropriate Hammett substitution parameter. This formalism expresses both sides of the equation in terms of free energy. An alternative formulation of this equation uses the parameter π which is the difference between the log*P* for the compound and the analogous hydrogen-substituted compound:

$$\log\left(\frac{1}{C}\right)=k_1\pi+k_2\sigma+k_3\,,\quad \text{where}\quad \pi=\log P_X-\log P_H \qquad (10.3)$$

Based on the shown examples, linear regression is the most widely used mathematical method to derive QSAR models. The simplest model is when only one dependent variable y with one independent variable x are in the equation: y = ax + b. In QSAR or QSPR y would be the property that one was trying to model (such as the biological activity) and x would be a molecular descriptor such as log*P* or a substituent constant [Leach2007].

To find values for the intercept *b* and slope *a* can be done by minimizing the sum of the squared differences between the values predicted by the equation and the actual observations:

LINK: http://www.youtube.com/watch?v=xojW6OEDfC4

For more than one independent variable, the method is referred to *as multiple linear regression* (the term *simple linear regression* applies where there is just one independent variable) [Leach2007].

The most common way to give the quality of the simple or multiple regression is calculating the squared correlation coefficient, or R^2 value which will be the determination coefficient. R^2 has a value between zero and one and it indicates the proportion of the variation in the dependent variable that is explained by the regression equation. R^2 can be calculated by defining Total Sum of Squares, TSS = $\sum_{i=1}^{N}(y_i-\bar{y})^2$, Explained Sum of

Squares, ESS = $\sum_{i=1}^{N}(y_{calc,i}-\bar{y})^2$, Residual Sum of Squares, RSS = $\sum_{i=1}^{N}(y_{calc,i}-y)^2$. R^2 is given by ESS/TSS = (TSS-RSS)/TSS = 1 - RSS/TSS, because TSS = ESS + RSS.

If the data (or the measurement error) have (multiple) normal distribution, R^2 of zero means that the variation in the observations is not at all explained by the variation in the independent variables; while R^2 of one means the perfect explanation. However in other data (or error) distribution, R^2 statistic can be misleading, because correlation and linearity will be not the same entity:

LINK: http://en.wikipedia.org/wiki/Correlation_and_dependence

Cross-validation methods provide a way to try and overcome some of the problems inherent in the use of the R^2 value alone [Leach2007]. Cross-validation involves the removal of some of the values from the data set, the derivation of a QSAR model using the remaining data, and then the application of this model to predict the values of the data that have been removed [Leach2007]. The simplest form of cross-validation is the leave-one-out approach (LOO), where just a single data value is removed [Leach2007]. Repeating this process for every value in the data set leads to a cross-validated R^2 (more commonly written Q^2 or q^2):

$$Q^2=1-\frac{PRESS}{T\sum_{i=1}^{N}(y_i-\bar{y}_{-1})^2}, \tag{10.4}$$

where PRESS is the Predictive Residual Sum of Squares which is another measure of predictive ability: $PRESS=\sum_{i=1}^{N}(y_{pred,i}-y_i)^2$¿. In PRESS instead of $y_{calc,i}$ used in RSS, the predicted values $_{pred,i}$ is used, which values are for data not used to derive the model. should strictly be calculated as the mean of the values for the appropriate cross-validation group rather than the mean for the entire data set [Leach2007].

Q^2 value is normally lower than the simple R^2. If there is a large discrepancy then it is likely that the data has been over-fit and the predictive ability of the equation will be suspect. A more rigorous procedure is to use an external set of molecules that is not used to build the model [Leach2007].

3. Introduction to the Advanced Statistical Methods

For many compounds and many descriptors the property matrix **X** can be defined:

$$\underset{N,M}{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix} \tag{10.5}$$

where N is the number of objects (e.g., compounds) and M is the number of variables (e.g., descriptors). Since the columns or the rows of the property matrix **X** can be correlated, several redundant information appears in the matrix **X**. Principal Component Analysis (PCA) can be transformed the original data into an abstract one which has orthogonal (uncorrelated) abstract variables:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathbf{T}}, \tag{10.6}$$

where matrix **T** will be the score matrix, and matrix **P** will be the loading matrix. The following video shows and explains visually the brief theoretical and practical background:

LINK: http://www.youtube.com/watch?v=UUxIXU_Ob6E&feature=iv&annotation_id=annotation_766703

The principal components (PCs) can be considered as a new orthogonal coordinate system, the projection of the original data matrix **X** to this new axes can be given by the following equation:

$$\mathbf{T}=\mathbf{XP}\,. \tag{10.7}$$

The new coordinates will be the linear combination of the original variables, e.g., for the elements of the first PC can be given as

$$\begin{aligned} t_{11}&=x_{11}p_{11}+x_{12}p_{21}+\ldots+x_{1M}p_{M1}\\ t_{21}&=x_{21}p_{11}+x_{22}p_{21}+\ldots+x_{2M}p_{M1}\\ &\ldots\\ t_{N1}&=x_{N1}p_{11}+x_{N2}p_{21}+\ldots+x_{NM}p_{M1} \end{aligned} \tag{10.8}$$

In principal components regression (PCR) the principal components are themselves used as variables in a multiple linear regression [Leach2007]. As most data sets provide many fewer "significant" principal components than variables (e.g. principal components whose eigenvalues are greater than one) this may often lead to a concise QSAR equation of the form:

$$y=b_1\mathbf{t}_1+b_2\mathbf{t}_2+b_3\mathbf{t}_3+\ldots \tag{10.9}$$

The following video shows and explains PCR in short:

LINK: http://www.youtube.com/watch?v=-5nnciZ9hgc

It is very important to mention, that the most important PCs is not necessarily will be involved in PCR, because PCA gives solution selecting PCs according to their ability to explain the variance in the independent variables whereas PCR is concerned with explaining the variation in the dependent y variable. One drawback of PCR is that it may be more difficult to interpret the resulting equations, for example to decide which of the original molecular properties should be changed in order to enhance the activity [Leach2007].

The technique of partial least squares regression (PLSR) is similar to PCR, with the essential difference that the quantities calculated are chosen to explain not only the variation in the independent variables x but also the variation in the dependent variables y as well. PLSR expresses the dependent variable in terms of quantities called latent variables which are linear combinations of the independent variables.

The following video shows and explains PLSR in short:

LINK: http://www.youtube.com/watch?v=WKEGhyFx0Dg

Because both PCR and PLSR use reduced dimensions, these methods give biased results, this is the price for treating the multicollinearity in the data. That is why the crucial step for both PCR and PLSR is to determine the proper number of PCs and latent variables, resp. Bootstrapping, Monte Carlo methods can help for that.

4. CoMFA (Comparative Molecular Field Analysis)

One of the most significant developments in QSAR in recent years was the introduction of Comparative Molecular Field Analysis (CoMFA). The aim of CoMFA is to derive a correlation between the biological activity of a series of molecules and their 3D shape, electrostatic and hydrogen-bonding characteristics [Leach2007]. The data structure used in a CoMFA analysis is derived from a series of superimposed conformations, one for each molecule in the data set. These conformations are presumed to be the biologically active structures, overlaid in their common binding mode. Each conformation is taken in turn, and the molecular fields around it are calculated. This is achieved by placing the structure within a regular lattice and calculating the interaction energy between the molecule and a series of probe groups placed at each lattice point. (In lines the compounds can be described, in every three columns theconformational Descartes coordinates can be found.) The general form of the equation is described in the next

$$activity = C + \sum_{i=1}^{N} \sum_{j=1}^{P} c(i,j) S(i,j) \quad (10.10)$$

where P probe groups, N points in the grid. c(i,j) is the coefficient for the column in the matrixthat corresponds group j at grid point i [Leach2007]. The solution of the PCA and PLS equation predict models for the 3D QSAR. Further trial based on the old methods are under development.

5. References

1. [Varnek2011] A. Varnek, I. I. Baskin, *Mol. Inf.* 2011, 30, 20-32.

2. [Baskin2008] I. I. Baskin, A. Varnek, Chapter 1. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening, 1-43. *in* A. Varnek, A. Tropsha (eds.), Chemoinformatics Approaches to Virtual Screening, Royal Society of Chemistry, Cambridge, 2008.

3. [Hansch1963] C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger and M. Streich, *J. Am. Chem. Soc.*, 1963, 85, 2817-2824.

4. [Hansch1964] C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, 86, 1616–1626.

5. [Free1964] S. M. Free Jr. and J. W. Wilson, *J. Med. Chem.*, 1964, 7, 395-399.

6. [Fleischer2000] R. Fleischer, P. Frohberg, A. Büge, P. Nuhn and M. Wiese, *Quant. Struct.-Act. Relat.*, 2000, 19, 162-172.

7. [Hatrik1996] S. Hatrik and P. Zahradnik, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 992-995.

8. [Leach2007] A. R. Leach, V. J. Gillet, An Introduction to Chemoinformatics. Springer, Dordrecht, The Netherlands, 2007.

6. Questions

1. How can you define chem(o)informatics?

2. What is the benefits of the linear relationship?

3. What is the benefits and drawbacks of the biased regression methods (PCR, PLSR)?

4. List some molecular descritptions?

7. Glossary

LINK: http://www.genomicglossaries.com/content/chemoinformatics_gloss.asp

Chapter 11. Quantum Mechanics and Mixed Quantum Mechanics/Molecular Mechanics Methods to Characterize the Structure and Reactions of Biologically Active Molecules.

(Gábor Paragi, György Ferenczy)

Keywords: Born-Oppenheimer approximation, potential energy surface, Hartree-Fock method, density functional theory, mixed methods, linked atom method, strictly localized molecular orbital method.

1. Introduction

What is described here? The aim of the chapter is to provide an introduction into the theoretical background of the most commonly used high level energy calculation methods. We review the different step in the simplification of the original problem, and the limitation of the most commonly applied calculation methods. Finally, we overview the theories as QM and MM level energy calculations can connect to each other within a large common system.

What is it used for? High level energy calculations in atomic level theoretical investigations.

What is needed? Beginner level knowledge in quantum mechanics.

2. The hierarchy of approximations in quantummechanical treatment of atoms and molecules.

Currently, the most complete theoretical description of atoms and molecules can be achieved by quantum mechanics (QM). Therefore its application in molecular biology seems an obvious step but the average size of biological systems as well as their complexity strongly limit the applicability of QM in biology. Introducing appropriate approximations in QM-level calculations, however, can help us to establish QM in the field of biology. Consequently, knowing the principles of applied approximations has primary importance in the relevancy or validity of the results at this level of investigations.

According to the general picture, quantum mechanics is the physics of "small" systems. The meaning of small is quite relative and obscure but more or less we can be assured that molecular or sub-molecular systems are "small" enough. Similarly to other part of physics (e.g. classical mechanics, electrodynamics, thermodynamics etc.) QM can also be built up in an axiomatic manner. For the curious reader we would refer to the related chapters in [1] but we do not wish to go into the details in the frame of the present lectures. According to these axioms, the QM level description of any state of a system is characterized by a wave function which is determined by the Schrödinger-equation (in non-relativistic cases). Until this point we only talked about systems in general but we have not defined clearly the elementary building blocks of our systems.

Many biochemical investigations focus onto the profound understanding of interactions between molecules which usually include bond creation or breaking. These processes are related to the changes in the electron system therefore the smallest elements are the electrons and nuclei and we investigate their systems called as atoms and molecules. It is known from introductory QM books (e.g. H_2^+ molecule), therefore in the investigation of biologically relevant systems several approximations must be applied. The detailed derivation of the approximations or the discussion about possible further developments would take a whole book and principally belongs to the field of atomic and molecular physics or theoretical chemistry. Therefore in the present chapter

we only would like to summarize shortly the most frequently applied approximations and show their validity region. We believe that for mainly application-oriented people or for MS students this is a good starting point.

3. From time-dependent systems to potential energy surface

3.1. The time-independent Schrödinger equation

As mentioned before, the evolution of a QM system in time is determined by the Schrödinger-equation in non-relativistic approximation. In case of molecules where the coordinates of the nuclei are signed by R_a (X_1, Y_1, Z_1, X_2, Y_2, Z_2, … , X_M, Y_M, Z_M; M = number of nuclei) and the electron coordinates by r_j (x_1, y_1, z_1, x_2, y_2, z_2, … , x_N, y_N, z_N; N = number of electrons) the Schrödinger-equation has the following form:

$$i\hbar\frac{\partial}{\partial t}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)=\sum_a\left(\frac{-\hbar^2}{2M_a}\Delta_{R_a}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\right)+\sum_j\left(\frac{-\hbar^2}{2m_e}\Delta_{r_j}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\right)+ \\ +V(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t) \quad (11.1)$$

Here "*i*" means the imaginary unit, ħ is the reduced Planck-constant, $V(\mathbf{R}_a,\mathbf{r}j\ ,t)$ signs any potential, Δ is the Laplace differential expression and finally $\Theta((\mathbf{R}_a,\mathbf{r}_j\ ,t)$ is the total wave function of the system. Ma means the mass of a nucleus while me is the mass of the electron. Hereafter, the "atomic units" will be applied in the whole chapter. It means that certain physical constants are chosen to unity, namely: ħ, m_e, a_0 (the Bohr-radius) and q_e (the electron charge). The energy unit is the Hartree and the exchange rate to other common energy units is the following: 1 Hartree = 0.5 Rydberg = 27.5 eV = 627.5 kcla/mol. Using atomic units, Eq. (11.1) turns to

$$i\frac{\partial}{\partial t}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)=\sum_a\left[-\frac{\Delta_{R_a}}{2M_a}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\right]+\sum_j\left(-\frac{1}{2}\Delta_{r_j}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\right)+V(\boldsymbol{R}_a \quad 11.2)$$

which is a simpler form of the same equation. According to the QM postulates, the first two terms on the right side is related to the kinetic energy of the nuclei and the electrons, respectively. The $V(\mathbf{R}_a,\mathbf{r}_j\ ,t)$ potential is built up from the electrostatic potential of the nuclei and the electrons, and any further external potentials (e.g. external electrostatic or magnetic field) can appear here as an extra additive term.

(Supplementary material) In QM we use different mathematical objects compared to the usual classical physics, so we would like to add a few words separately about them. If somebody is familiar with the basics of linear algebra, he/she can easily jump this supplementary part.

Taking a set of selected real or complex value functions (e.g. set of solutions of eqn. (1)) one can define linear-space or vector-space structure on this set with the help of usual scalar multiplication and addition of functions. In this situation the elements of the set are called generally vectors. Those objects which map between two vector spaces (or maps a vector space onto itself) with certain mathematical properties are called linear operators, or simply operators. There is a special situation when the image set of the mapping is the real numbers. In this case the operator called as functional and later we will work with this mathematical object in the frame of density functional theory. A good example for a functional is the definite integral: it assigns the area under the curve value of the chosen domain to the function in the integral. For a curious reader, a detailed introduction into the mathematics of vector spaces can be found in the book of P. R. Halmos [2].

Consequently the $V(\mathbf{R}_a,\mathbf{r}_j\ ,t)$ potential or the Laplace expression is an operator. The $V(\mathbf{R}_a,\mathbf{r}_j\ ,t)$ operator assigns to a wave function the multiple of the function according to $V(\mathbf{R}_a,\mathbf{r}_j\ ,t)\Theta(\mathbf{R}_a,\mathbf{r}_j\ ,t)$. The Laplace-operator is a little bit more complicated: it is evaluated as the sum of second order partial derivatives with respect to the variables. For instance, the Laplace-operator of the helium atom is as follows: We have one nucleus and two electrons, therefore the independent variables of the wave function are (X_1, Y_1, Z_1, x_1, y_1, z_1, x_2, y_2, z_2) in Descartes-coordinates. Thus the Laplace-expression is

$$\Delta\Theta=\frac{\partial^2\Theta}{\partial X_1^2}+\frac{\partial^2\Theta}{\partial Y_1^2}+\frac{\partial^2\Theta}{\partial x_1^2}+\frac{\partial^2\Theta}{\partial y_1^2}+\frac{\partial^2\Theta}{\partial z_1^2}+\frac{\partial^2\Theta}{\partial x_2^2}+\frac{\partial^2\Theta}{\partial y_2^2}+\frac{\partial^2\Theta}{\partial z_2^2} \quad \text{hence}$$
$$\Delta=\frac{\partial^2}{\partial X_1^2}+\frac{\partial^2}{\partial Y_1^2}+\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial z_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial z_2^2}=\sum_{a=1}^{M=1}\Delta_{R_a}+\sum_{i=1}^{N=2}\Delta_{r_i} \qquad (11.3)$$

in Descartes-coordinates. It is worth to note that according to the axioms of QM the physical quantities are described by operators contrary to the classical physics, where they are usually real value functions. Hereafter we will use the word "operator" together with physical quantities (e.g.impulse-operator, coordinate-operator etc.), and the hat (^) above a letter denotes that it is an operator.

As an example, the impulse-operator of a particle is $\hat{p}=-i\hat{\nabla}$ and the kinetic energy of a particle is evaluated as $p^2/2m$. The corresponding operator is $\hat{T}=(-i\hat{\nabla})^2/2m=-\Delta/2m$ and defining

$$\hat{T}=\hat{T}_{nuclei}+\hat{T}_{electron}\equiv\hat{T}_n+\hat{T}_e \;;\; \text{where} \quad \hat{T}_{nuclei}\equiv\hat{T}_n=\sum_{a=1}^{M}-\frac{\Delta_{R_a}}{2M_a} \;;\; \hat{T}_{electron}\equiv\hat{T}_e=\sum_{j=1}^{N}-\frac{\Delta_{r_j}}{2} \qquad (11.4)$$

one can write the Schrödinger equation as

$$i\frac{\partial}{\partial t}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)=\hat{T}\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t)+\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j,t)\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j,t) \qquad (11.5)$$

(**end of supplementary material**)

If the potential does not depend explicitly on time - so it has the form $V(R_a,r_j)$ - then we can look for the wave function solution of the Schrödinger equation in $\boldsymbol{\theta}(\mathbf{R}_a,\mathbf{r}_j,t)=\Phi(\mathbf{R}_a,\mathbf{r}_j)\chi(t)$ form. This method is called as separation of variables in mathematics since the original many variable function is written as the product of functions with fewer variables than the original one. The new functions still can have many variables but cannot have the same one simultaneously. The advantage of this method is that the original equation with large number of variables is separated into more than one independent equation but with fewer variables.

In the present case the original wave function separated into a purely time-dependent function and a function with all the other variables. Substituting the new form into Eq. 11.1 and take the advantage of that the purely time dependent function behaves as a constant function for the kinetic energy operator, after some algebra we can have

$$\frac{i}{\chi(t)}\frac{d\chi(t)}{dt}=\frac{1}{\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)}(\hat{T}\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)+\hat{V}(\boldsymbol{r},\boldsymbol{R}_a,\boldsymbol{r}_j)\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)) \qquad (11.6)$$

The left side of Eq. 11.2 depends only on the time variable while the other side on all the other ones, which can only happen if both sides are constant. Let's denote this constant first with ω (or in non-atomic units with ħω) and we get $\chi(t)=A\cdot\exp(-i\cdot\omega t)$ by solving the equation from the left side. This way the general form of the purely time-dependent part of the wave function has been determined.

Regarding the right-side of Eq. 11.2 , if $\Phi(\mathbf{R}_a,\mathbf{r}_j)$ is moved from the denominator next to the ω constant, and the constant will be denoted hereafter as E (which is understandable if we take into account the non-atomic units form of the constant ħω), then we can get the Schrödinger equation of stationary states, namely

$$E\cdot\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)=\hat{T}\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)+\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j)\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j)=\hat{H}\Phi(\boldsymbol{R}_a,\boldsymbol{r}_j), \quad \text{where} \quad \hat{H} \qquad (11.7)$$

The newly defined operator $\hat{H}$ is called as the Hamilton-operator of the system, and E is the total energy of the system. Since Eq. 11.3 does not depend explicitly on time therefore $\Phi(\mathbf{R}_a,\mathbf{r}_j)$ will not change in time as well. These solutions are known as stationary states, and the next step is the solution of this equation.

It is worth to mention that those equations where we try to find those vectors, on which the effect of an operator is at most a scalar multiplication, are known as eigenvalue equations. The solutions of such an equation are called eigenfunctions (or eigenvectors), and the scalars are the eigenvalues. This is the case in Eq. 11.3 if one looks the left and right parts of the formula. The $\hat{H}$ operator acts on the $\Phi(\mathbf{R}_a,\mathbf{r}_j)$ function and provide its scalar multiplied form ($E\cdot\Phi(\mathbf{R}_a,\mathbf{r}_j)$).

3.2. The adiabatic and the Born-Oppenheimer approximations

The determination of the stationary states can be achieved by the solution of Eq. 11.3, however, it is still a complicated problem because of the large number of variables. Thus the primary aim of the further approximations is still to reduce the number of variables. Taking into account that from biological and chemical point of view the most important effects (bond breaking or creation) are related to the electron system, we try to simplify Eq. 11.3 that way that the variables of nuclei will be eliminated. Since the total mass of the nuclei is 3-4 orders of magnitude larger than the total mass of the electron system, the electrons can instantaneously follow the motion of nuclei. If we ignore the backward coupling of the electron system to the nuclei, the motion of nuclei and the motion of electrons are considered independently. Thus the coordinates of the nuclei can be considered as parameters for the electron coordinate in a chosen instant. Fixing the geometry of nuclei we can get the eigenfunction of the electron system related to the chosen nucleus-geometry. Introducing the notation $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$ and E_k for the k-th eigenfunction and eigenvalue of the electron system in the fixed nuclei, where the lowest eigenvalue belongs to k=1. It can be shown that the $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$ eigenfunctions form a basis in the vector-space of all solutions related to any fixed geometry. Mathematically it means

$$\Theta(\boldsymbol{R}_a,\boldsymbol{r}_j)=\sum_k \chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j) \tag{11.8}$$

The $\chi_k(\mathbf{R}_k)$ coefficients certainly depends on the fixed geometry of the nuclei, therefore they contain the nuclei coordinates as parameters. If we use the decomposition of the $\hat{T}$ operator based on the separation of the nuclei and electron coordinates ($\hat{T}=\hat{T}_{nuc}+\hat{T}_{elec}$) and introduce the Eq. 11.5

$$\hat{H}_{elec}=\hat{T}_{elec}+\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j) \quad \text{with eigenvalue equation} \quad \hat{H}_{elec}\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)=E_k\Psi_k \tag{11.9}$$

called as the Hamilton operator of the electron system, then the form of Eq. 11.3 will change to

$$\begin{aligned} E\cdot\sum_k \chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)&=(\hat{T}_{nuc}+\hat{H}_{elec})\sum_k \chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)= \\ &=\sum_k(\hat{T}_{nuc}(\chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j))+\chi_k(\boldsymbol{R}_a)\hat{H}_{elec}\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)) \\ &=\sum_k(\hat{T}_{nuc}(\chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j))+(\chi_k(\boldsymbol{R}_a)E_k\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)) \end{aligned} \tag{11.10}$$

$$E\cdot\sum_k \chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)=\sum_k(\hat{T}_{nuc}(\chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j))+\chi_k(\boldsymbol{R}_a)E_k\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j$$

by substituting Eq. 11.4 and Eq. 11.5 into Eq. 11.3. Here we applied the property of addition of operators then we also used that the $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$ is an eigenfunction of $\hat{H}_{elec}$ with E_k eigenvalue. Acting $\hat{T}_{nuc}$ onto the $\chi_k(\mathbf{R}_a)$ $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$ product – without the details of the calculation – we get for the first term on the right side in Eq. 11.6

$$\hat{T}_{nuc}(\chi_k(\boldsymbol{R}_a)\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j))=\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)(\hat{T}_{nuc}\chi_k(\boldsymbol{R}_a))+(\hat{B}\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j))\chi_k(\boldsymbol{R}_a) \tag{11.11}$$

where $\hat{B}$ the new operator contains the first and second derivatives with respect to the variables of nuclei. Let's substitute Eq. 11.7 into Eq. 11.6 then multiply both side with $\Psi^*_l(\mathbf{R}_a,\mathbf{r}_j)$ (the star means the complex conjugation for complex value wavefunction) and integrate according to the variables of the electron. This way we can get matrix element of an operator (e.g.: $\hat{B}$) with respect to the chosen basis set (in our case $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$). Hereafter the

(l,k)-th matrix element of an operator will be denoted by the indexes in the upper-right position of the operator, so

$$\hat{B}^{lk}(\boldsymbol{R}_a)=\int \Psi_l^*(\boldsymbol{R}_a,\boldsymbol{r}_j)\hat{B}\Psi_k(\boldsymbol{R}_a,\boldsymbol{r}_j)d^3\boldsymbol{r}_1\cdot d^3\boldsymbol{r}_2\cdot\ldots\cdot d^3\boldsymbol{r}_N \qquad (11.12)$$

By applying the orthonormality of the basis functions $\Psi_k(\mathbf{R}_a,\mathbf{r}_j)$, the form of Eq. 11.6 turns into

$$E\cdot\chi_l(\boldsymbol{R}_a)=\hat{T}_{mc}\chi_l(\boldsymbol{R}_a)+\chi_l(\boldsymbol{R}_a)E_l+\sum_k \hat{B}^{lk}(\boldsymbol{R}_a)\chi_k(\boldsymbol{R}_a) \quad (l=1,\ldots,M_a) \qquad (11.13)$$

with the help of Eq. 11.7 and Eq. 11.8. In the system of Eq. 11.9 the $\chi_l(\mathbf{R}_a)$ functions and the E are the unknown quantities and the equations form a coupled system of equations because of the last term in Eq. 11.9. The E_l quantities can be determined by solving the eigenvalue equation in Eq. 11.5. If we set the $k \neq l$ matrix elements of $\hat{B}$ to zero, then we decouple the equations in Eq. 11.9 so we will have M_a independent equations. This approximation, when the $\hat{B}^{kl}$ elements are set to zero called as **adiabatic approximation**, and it should be distinguished from the **Born-Oppenheimer approximation**, where the whole $\hat{B}$ matrix is set to zero. It means that in the Born-Oppenheimer approximation both the diagonal and the off-diagonal elements of the $\hat{B}$ operator are set to zero. Thus the principle question is now the solvation of eigenvalue equation in Eq. 11.5. Before we start to focus onto this problem, we would like to have a few words on the consequences of the Born-Oppenheimer approximation.

3.3. The potential energy surface

Let's write Eq. 11.9 in the Born-Oppenheimer approximation as

$$E\cdot\chi_l(\boldsymbol{R}_a)=\hat{T}_{mc}\chi_l(\boldsymbol{R}_a)+\chi_l(\boldsymbol{R}_a)E_l=(\hat{T}_{mc}+E_l)\cdot\chi_l(\boldsymbol{R}_a) \quad (l=1,\ldots,M_a) \qquad (11.14)$$

The form of the equation is strongly reminds of the form of a Hamilton-operator, where the nuclei are handled independently from the electrons and they move in the "potential" denoted by E_l . Therefore the elimination of the term with the $\hat{B}$ operator really means the full decoupling of the motion of nuclei from the electron system. Moreover, the El depends implicitly on the geometry of the nuclei, so changing the geometry of the nuclei (i.e. moving them) provide different El values. Plotting the El values as the function of the geometrical parameters, we can get a "surface" in the parametrical coordinate system. This surface called as potential energy surface (PES). Taking into account the principle of minimum total potential energy, the nuclei prefer those geometries where the El surface has its minima. This is the theoretical basement of geometrical optimization calculations, where the total energy of the electron system calculated in different arrangement of the nuclei. It is worth to note that the state of the electron system (e.g. ground state or any excited state) cannot change on a surface: different states of the electrons provide different PESs.

As an example, we present the ground state PES of the ethanol molecule with respect to two dihedral parameters (see figure 11-1). The two parameters describe the rotation of the methyl and the hydroxyl groups around the C-C and the C-O bonds, respectively. The ground state energies of the electron system are calculated systematically in 30° steps. The different minima and maxima of the surface can be easily interpreted by staggered and eclipse geometries of the rotating groups.

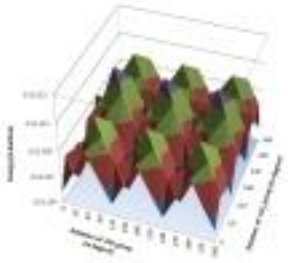


Figure 11.1. The PES of the ethanol regarding the systematic rotation of the methyl and the hydroxyl group.

We would like to note that in a later chapter a case study will show the application of PES in a conformational analysis of a biologically interesting molecule.

4. Solving the Schrödinger equation of the stationary N-electron system

Let's have a quick summary about the approximations which have been applied until this point in the QM level discussion of atoms and molecules. First, non-relativistic situation has been chosen since we use the Schrödinger equation. Then we separated the time coordinate and considering only the stationary states following by the separation of the nucleus and the electron coordinates with the help of the adiabatic and the Born-Oppenheimer approximation. To determine the ground or excited states energies (and wave functions) of the electron system we have to solve the eigenvalue equation of the $\hat{H}_{elec}=\hat{T}_{elec}+\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j)$ operator which is still not an easy task. It is well known that system with more than 2 electrons cannot be solved analytically, and one can imagine that biologically or chemically interesting systems easily contains several hundreds of electrons. So, further approximations or omissions are still surely necessary.

Before having the next step, we have to clarify an important fact regarding identical particles. Two particles are considered as identical, if all of their inner properties (mass, charge, spin, etc.) are the same. In classical mechanics it is not problematic to distinguish two identical particles since both of them have a well determined path. However, this is not the case in quantum mechanics. If we have a QM system with identical particles they are indistinguishable and this requires the introduction of a new postulate. Namely, we require that the wave function of a system must be completely symmetric if it is built up from identical particles with integer spin (bosons). On the other hand, the wave function of a system must be completely antisymmetric if it contains identical particles with half spin (fermions). The antisymmetric property means that the wave function of the system must change its sign for the exchange of two particles. Since the electron has half spin therefore the wave function of the electron system must be totally antisymmetric.

Until this point we did not specify the concrete form of the potential term in the $\hat{H}_{elec}$ operator. If we do not have any additional external fields (e.g. extra electrostatic or magnetic field) then the $\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j)$ potential contains three terms based on electrostatic interactions: the repulsion of the nuclei, the attraction of the nuclei and the electrons and the repulsion of the electrons. More concretely,

$$\hat{V}(\boldsymbol{R}_a,\boldsymbol{r}_j)=\hat{V}_{nn}(\boldsymbol{R}_a,\boldsymbol{R}_b)+\hat{V}_{ne}(\boldsymbol{R}_a,\boldsymbol{r}_j)+\hat{V}_{ee}(\boldsymbol{r}_j,\boldsymbol{r}_k)= \\ =\sum_{a<b}\sum_{b}\frac{Z_aZ_b}{|R_a-R_b|}-\sum_{j}\sum_{a}\frac{Z_a}{|R_a-r_j|}+\sum_{j<k}\sum_{k}\frac{1}{|r_j-r_k|} \tag{11.15}$$

where Z_a means the total charge of the nucleus in the position of R_a. It is important to note that the kinetic energy operator of the electron system has the form

$$\hat{T}_{elec}=\sum_{j}-\frac{\Delta_j}{2}=\sum_{j}\hat{t}_j \tag{11.16}$$

Taking these forms, the terms in the $\hat{H}_{elec}$ operator can be classify according to the sub-terms in the summations in Eq. 11.11 and Eq. 11.12. The classification is based on the number of electrons appear through variables in a sub-term. Regarding the potential operator, the first term does not contain any electron variable, it contains the coordinates of one electron in each sub-term and the last term contains the coordinates of two particles in every sub-terms. The kinetic energy operator contains the variable of one particle in each sub-terms in Eq. 11.12. Aside from the nucleus-nucleus interaction term, we have two classes: the one-particle and the two-particle operators. Thus, the kinetic and the nucleus-electron interaction operators are one-particle operators, while the electron-electron interaction operator is a two-particle operator. The nucleus-nucleus interaction operator

behaves as a constant multiplier regarding the wave function of the electron system and in the eigenvalue-equation it shifts with a constant the Ej values so it will be taken into account only at the end of the calculations.

Focusing onto the eigenvalue equation of the electrons, such methods will be reviewed first briefly which replace the two-particle operator in the original potential with an effective one-particle potential. This way the electrons will move in a potential created by the nuclei and the new effective potential without having interaction with each other. It can be also interpreted as every electron in the system would move independently, and the effect of the other electrons would be taken into account in an averaged manner by the effective potential. That's why these approaches usually called as independent particle approximation or mean-field approximation. The primary consequence of these methods is that the original equation with $3N_e$ (or by taking into account the spin, with 4 N_e) variables is substituted by N_e equations with 3 (or 4) variables. So, the Hamilton operator of the electron system without the nucleus-nucleus interaction can be transformed into the sum of one-particle operators. Hence we can write

$$\hat{H}_{elec} - \hat{V}_{nn} \approx \sum_j \hat{f}_j \quad , \quad \text{where} \quad \hat{f}_j \equiv \hat{f}(r_j) = \hat{t}(r_j) + \hat{v}_{ne}(r_j) + \hat{v}_{eff}(r_j) \tag{11.17}$$

but the explicit form of $\hat{v}_{eff}(r_j)$ is still unknown. If we can determine the $\hat{v}_{eff}(r_j)$ term then the variables can be separated in the eigenfunction by applying a product form. Here the terms in the product are the one-particle functions which depends on the 3 (or 4) variables of one electron. Moreover, this is the first step to include the orbital picture into the QM level treatment of the electron system. However we cannot forget about the antisymmetric property of the wave function regarding the electron system. Therefore we cannot build up the total N-particle wave function as a *simple* product of the one-particle functions, since it must satisfy the antisymmetric requirement. Considering the simplest two-particles system, the form

$$\Psi(r_1, r_2) = \frac{1}{\sqrt{2}}(\varphi_1(r_1)\varphi_2(r_2) - \varphi_1(r_2)\varphi_2(r_1)) \tag{11.18}$$

is the simplest choice for an antisymmetric expression, where the factor in the beginning of the expression is the consequence of the normalization. One can easily check that if we exchange the two variables, then the expression change the sign. Similarly, in case of three electrons a 6-members product provide the same results, as well as a 24-members product in case of four particles. John C. Slater was the first, who derived the antisymmetrical total wave function with the help of matrix determinant. Namely, he wrote the total wave function as the result of the calculation of the determinant of the following N_exN_e matrix

$$\Psi(r_1, \ldots, r_{N_e}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(r_1) & \cdots & \varphi_1(r_{N_e}) \\ \vdots & \ddots & \vdots \\ \varphi_{N_e}(r_1) & \cdots & \varphi_{N_e}(r_{N_e}) \end{vmatrix} \tag{11.19}$$

The vertical lines refer to the determinant calculation, and one can easily check that this is the general and compact form of the previously mentioned 2, 3 or 4 particles cases. It is worth to note that the determinant form is not the only one which can provide a totally antisymmetric wave function based on the product of one-particle functions. However this is one of the simplest ones, and the question now is only that, how can we determine the unknown effective potential? The firs answer for this problem was given by the Hartree-Fock method.

4.1. The Hartree-Fock method

In the derivation of the Hartree-Fock equations (or herinafter HF) we utilize that the ground state energy of the system is an extremum. Consequently, the HF scheme is not a general method since it can approximate only the ground state energy and the ground state wave function. Moreover, the solutions are restrained by the condition that the ground state wave function must derive from a one-determinant expression where the one-particle functions form an orthonormal system. Therefore the HF method is usually known as the one-determinant approximation, and this is an alternative formulation of the independent particle picture.

Without going into the details, we would like to summarize the derivation of the HF-equations. First, expressing the ground state energy of the system with the help of the eigenvalue-equation as

$$\frac{\int \Psi^*(r_j)(\hat{H}_{elec}-\hat{V}_{nn})\Psi(r_j)d^3r_1\ldots d^3_{N_e}}{\int \Psi^*(r_j)\Psi(r_j)d^3r_1\ldots d^3_{N_e}}=E \qquad (11.20)$$

If the ground state wave function $\Psi(\mathbf{r}_j)$ is a one-determinant expression, then we can substitute different one-particle trial functions into the determinant and we can think about Eq. 11.16 as if it were the functional of the one-particle functions, so $E= E(\varphi_1 \ldots \varphi_{N_e})$. Moreover, it is also true that this functional reaches its minimum when the ground state wave function is substituted into the formula, and in this case the value of E is exactly the ground state total energy. Therefore the mathematical task is the minimization of the energy functional with a suitable set of (orthonormal) one-particle wave function. The necessary condition for the extremum is that the first variation of the expression must vanish. This condition can help us to determine the one-particle operator, whose eigenfunctions of the lowest N_e eigenvalues provide the best one-particle function set. Writing here only the final results of the conditional variation

$$\hat{f}(r_j)\varphi_\lambda(r_j)=\varepsilon_\lambda\varphi_\lambda(r_j) \quad \text{where}$$

$$\hat{f}(r_j)\varphi_\lambda(r_j)=\left[-\frac{\Delta_j}{2}\right]\varphi_\lambda(r_j)-\left[\sum_a^{N_n}\frac{Z_a}{|r_j-M_a|}\right]\varphi_\lambda(r_j)+ \qquad (11.21)$$

$$+\left[\sum_\mu^{N_e}\int\frac{\varphi^*_\mu(r_k)\varphi_\mu(r_k)}{|r_j-r_k|}d^3r_k\right]\varphi_\lambda(r_j)-\sum_\mu^{N_e}\int\frac{\varphi^*_\mu(r_k)\varphi_\lambda(r_k)}{|r_j-r_k|}d^3r_k\varphi_\mu$$

these one-particle equations must be solved. Taking the eigenfunctions solution of the lowest N_e eigenvalues, we can build up the one-determinant matrix and calculating the ground state wave function. Following this, the ground state energy can be evaluate with the help of Eq. 11.16. In Eq. 11.17 the last two terms is the HF form of the effective potential, where the first one is the coulomb term and the second one is the so-called exchange term.

Until this point the one-particle picture seems as a mathematical necessity in the reduction of variables. However with the help of the Koopman-theorem, a physical interpretation can be put behind the one-particle picture. Namely, the Koopman-theorem proves that those one-particle eigenvalues whose eigenfunction took part in the construction of the matrix (occupied orbitals), provide a good approximation for the ionization energies. Moreover, the eigenvalues of those orbitals which do not take part in the construction of the matrix (unoccupied orbitals) can be interpreted as an approximation of the electron affinity.

During the solution of Eq. 11.17, one can face with the problem, that the operator itself contains the unknown eigenfunctions. The problem can be handled by the application of the Self Consistent Field method, or briefly, SCF-method. This is an iterational method where the operator in the new step is built up from the eigenfunction of the previous step, and then solving the equation. These iteration steps are going on until that point while the changes in the chosen quantities (e.g. eigenvalue or potential) between two iteration steps do not decrease under a certain threshold. When the changes are smaller than the threshold we can say that the system converged, and the eigenvalues and the eigenfunction of the last iteration is considered as the solution of the eigenvalue equation.

4.2. The Density Functional Theory

Beside the HF or the HF-based methods another very popular method is the Density Functional Theory (DFT) for the determination of the wave function or the energy of the electron system. In its original form DFT was capable to handle only non-relativistic stationary ground state systems but from the 1980's this theory was gradually extended to further phenomena, like excited states, time-depending or relativistic events, and many more fields of interests [3]. However, here we discuss only the basics of the theorem.

In the theoretical background of DFT two important steps must be distinguished. First, two theorems are proved (Hohenber-Kohn theorem I and II [4]), and for the understanding of them we need some further definitions.

Let's take the potential term of $\hat{H}_{elec}$ according to Eq. 11.11 and fix the number of electrons (N_e). Then we can define two sets: i, taking the set of the external potentials $\{V_{ext}\}$ which originated from the different arrangements of the nuclei in the absence of any additional external field. ii, Let's have the set of the ground state total densities of the electron system $\{\rho_{gs}\}$ related to the different external potentials (i.e. to the different geometry of the system).

The first Hohenberg-Kohn theorem (HK-I) demonstrates that there is a bijective map between this two set. A simple consequence of the theorem is that if we know the ground state total density then all the properties of the system can be evaluated since the full Hamiltonian is determined according to Eq. 11.11. The most important consequence of the bijection is that we can define an energy functional, which now is the functional of the ground state density contrary to the HF method where the energy expression is the functional of the one-particle wavefunctions. Because of Eq. 11.11 the energy functional has the following terms

$$E[\rho]=T_{kin}[\rho]+E_{ext}[\rho]+E_{ee}[\rho], \text{ with the usual } E_{ee}[\rho]=E_{coulomb}[\rho]+\ldots \quad (11.22)$$

decomposition. Here $T_{kin}[\rho]$ is the kinetic energy functional of the electron system and $E_{ext}[\rho]$ is the electrostatic energy functional of the nuclei and all other possible external field potentials. $E_{ee}[\rho]$ denotes the electron-electron interaction energy functional, where $E_{coulomb}[\rho]$ means the electrostatic repulsion energy functional between the electrons and $E_{XC}[\rho]$ signs the remaining part of $E_{ee}[\rho]$. This latter statement is the official definition of the $E_{XC}[\rho]$ term, called as the exchange-correlation functional: Taking away certain known part from the unknown $E_{ee}[\rho]$ functional, and giving a name to the remaining unknown part. It is worth to note that the name of this unknown part refers to the fact that in HF theory this remaining part can be expressed with the one-particle functions, and called as exchange term (cf. last term in Eq. 11.17)

According to the second Hohenberg-Kohn theorem (HK-II) the $E[\rho]$ energy functional reaches its minimum if we substitute that ground state density which is assigned to the potential in the $E_{ext}[\rho]$ term by the HK-I theorem. Moreover, the minimum value of the functional in this case is exactly equal with the ground state energy of the system.

It is worth to not that until this point the ground state total energy of the electron system in principle can be determined exactly while in HF theory – because of the one-determinant approximation in the energy expression – one *ab ovo* cannot determine exactly the ground state energy of the system.

Moreover, in Eq. 11.18 only the $E_{ext}[\rho]$ and the $E_{coulomb}[\rho]$ functional forms are known and all the others are not. In the frame of the Thomas-Fermi theorem an approximate form is derived for the $T_{kin}[\rho]$ term in case of ground state systems. Few years later, P.A.M. Dirac successfully expressed the averaged HF exchange energy term as the functional of the total electron density [5] and augmented the original Thomas-Fermi method with this new expression. It is important to note that during the derivation of Dirac's expression the ground state property of the system was applied as well as in case of the kinetic energy expression derivation. Therefore, the Thomas-Fermi-Dirac theorem can determine only the ground state energy and density of the system similar to the HF method. Moreover, the Thomas-Fermi-Dirac theorem had been derived by heuristic considerations nearly 20 years before proving the Hohenberg-Kohn's theorems. Consequently, this model was also theoretically established by the two Hohenber-Kohn theorems.

The second important step in DFT is that the HK-theorems provide the basement of an alternative calculation scheme, which nowadays is the most popular application form of theorem. This is the Kohn-Sham (KS) picture [6] which is essentially the application of the independent particle picture in the frame of the DFT. More precisely, in the Kohn-Sham picture we substitute the original interacting system with a virtual one having the same number of non-interacting electrons. The connection between the original and the virtual system is that we require that the two systems have the same ground state density. The HK-I theorem ensures a bijection between the set of ground state densities (which is the same for the interacting and non-interacting systems because of the requirement) and the set of the effective potentials related to the virtual system. The HK-I theorem holds for this bijection, since in the independent particle picture the effective potential plays the role of the external potential. For the better understanding of the Kohn-Sham picture we summarized the basics of the theorem in *Figure 11.2*.

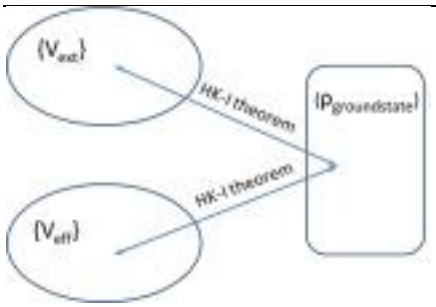


Figure 11.2. The schematic representation of the Kohn-Sham picture.

However until this point we do not know anything about the explicit form of the effective potential except that it is the sum of the - yet unknown - one-particle effective potentials. Without going into the details of the derivation of the Kohn-Sham equations we present the form of the one-particle equations in Eq. 11.19.

$$\left[\frac{-\Delta_j}{2}+v_{eff}^{KS}(r_j)\right]\varphi_\lambda^{KS}(r_j)=\varepsilon_\lambda^{KS}\varphi_\lambda^{KS}(r_j) \quad \text{where}$$

$$v_{eff}^{KS}=v_{ext}(r_j)+\int\frac{\rho(r_k)}{r_j-r_k}d^3r_k+v_{XC}(r_j) \quad \text{and} \quad \rho(r_j)=\sum_{\lambda=1}^{N_e}\varphi_\lambda^{KS}(r_j)^2 \qquad (11.23)$$

Here the first term in the effective Kohn-Sham potential is the usual external potential based on the geometrical arrangement of the nuclei. The second term comes from the electrostatic repulsion between the electrons. The third term is the unknown exchange-correlation potential which is the functional derivative of the exchange-correlation energy functional. Now it is understandable why so important in DFT calculations the proper choice of the exchangecorrelation potential: only this term has approximate form in the Kohn-Sham potential, so the accuracy of the calculation primarily determined by this term. In the solution of the KS equations again the SCF technique should be applied, since the KS potential contains the total density but the total density is built up from the KS-orbitals. It is worth to note again that while the HF-method in principle cannot determine the exact ground state energy, the DFT in principle is an exact theorem: if we would know the exact form of the exchange-correlation term then we could determine exactly the ground state density, and by this way the ground state energy. Finally we would like to mention that in the derivation of the KS equations we applied the ground state character of the electron system therefore the KS-equation holds for only ground state systems. Moreover, in many cases the ground state feature of the system was also used in the derivation of the exchange-correlation term, so DFT can be applied to only ground state systems in this form.

5. Rational for mixed QM/MM (QM/QM) methods

Quantum mechanics (QM) offers a potentially accurate description of chemical systems including their structure and energetics. In contrast to classical force fields QM is able to describe systems far from their equilibrium geometry and thus it can be applied to study chemical reactions. However, QM is computationally intensive and large systems like those typical in biochemical problems (namely biopolymers in aqueous environment) cannot be treated by routine high level QM methods at a reasonable computational effort.

Mixed QM/MM methods are based on the idea that many biochemical phenomena including biochemical reactions, structural changes and spectroscopic events can be described by applying a QM method to describe electronic changes localized to a certain region of the system while a more approximate method is appropriate for the rest of the system. As an example, let us quote enzymatic reactions where typically the substrate and few surrounding residues are directly involved in electronic changes while the rest of the system exerts its effect primarily by electrostatic interactions. Then it is advantageous to separate the total system into two parts. The central subsystem comprises the part where electronic changes take place and it is embedded in the larger outer subsystem or environment. The computational treatment of the enzymatic reaction can exploit this separation by performing a high level QM description of the central subsystem with a lower level description of the environment. The QM calculation for the central subsystem can account for the electronic changes and the

computationally less demanding method applied for the environment can cope with the more extended outer subsystem (*Figure 11.3*).

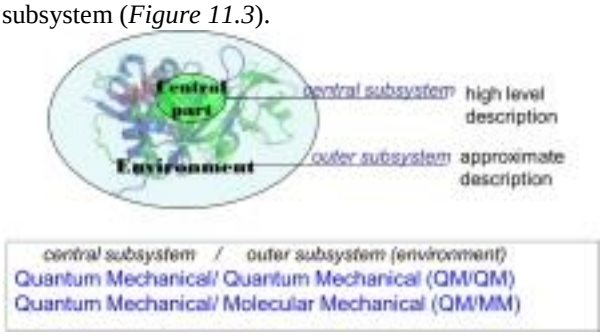


Figure 11.3. Separation of a large system into subsystems that are treated at different level computational methods

Mixed methods apply QM for the central subsystem and they may apply molecular mechanics (MM) for the outer subsystem. Such schemes are called QM/MM methods. The outer subsystem may also be treated with a lower level QM method and such schemes are called QM/QM methods. Further subdivision of the total system is also possible leading to for example QM/QM/MM methods. The following discussion presents QM/MM methods with occasional reference to QM/QM methods.

5.1. .Energy expressions in mixed methods

There are two main energy evaluating schemes used in QM/MM methods. Additive energy expressions include three terms

$$E_{QM/MM}^{C+E} = E_{QM}^{C} + E_{MM}^{E} + E_{coupling}^{C,E} \quad (11.24)$$

where E_{QM}^{C} is the energy of the central subsystem at QM level, E_{MM}^{E} is the energy of the environment at MM level and $E_{coupling}^{C,E}$ is a coupling term describing the interaction between the subsystems. Based on the way this latter term is evaluated three coupling schemes are distinguished. Mechanical embedding uses MM terms only in $E_{coupling}^{C,E}$. The form of the interaction terms agrees with that of the MM force field. In particular, electrostatic interactions are calculated with MM point charges assigned to the QM system and charges are not updated with changes in the wave function e.g in a chemical reaction. The next level is electrostatic embedding that calculates the interaction of MM charges with the wave function of the central subsystem. In this way, the wave function accommodates to the electrostatic changes in the environment. This is the most commonly applied coupling scheme owing to the significant improvement it represents over the mechanical coupling and also to its relatively easy implementation. It should be noted however, that the use of charges derived for an MM force field is not necessarily the optimal choice for describing interactions with the central part wave function. An even higher level of coupling called polarized embedding. This more sophisticated approach takes into account the change of the charges (and possibly higher moments) in the environment due to the field of the wave function. Various implementations of the polarized embedding have been proposed, but no common practice for applying polarization emerged so far. Reasons for this include that polarizable MM force fields are not routinely available. Furthermore, polarized MM charges (and higher moments) back-polarize the wave function and this mutual polarization calls for an iterative treatment.

Subtractive energy expressions require MM calculations for the whole system (E_{MM}^{C+E}) and for the central subsystem (E_{MM}^{C}) in addition to a QM calculation for central subsystem (E_{QM}^{C}). The total energy is written as

$$E_{QM/MM}^{C+E} = E_{MM}^{C+E} + E_{QM}^{C} - E_{MM}^{C} \quad (11.25)$$

That is, the MM energy of the whole system is improved by adding the difference of the QM and MM energies of the central subsystem. The advantage of this energy evaluating scheme is its simplicity. On the other hand, difficulties may arise in finding appropriate MM parameters for the central subsystem (e.g. in transition states of chemical reactions).

5.2. Subsystem separation

The separation of the total system into central part and environment is straightforward when there is no chemical bond between the subsystems. A possible example is the treatment of the chemical reaction of small molecules in water, where the QM treatment of the solute molecules together with few water molecules and the MM treatment of all other water molecules is a reasonable approach. On the other hand, QM and MM subsystems are inevitably connected by covalent bonds in most enzymatic reaction computations. Typically, protein residues participate in the electronic rearrangements and therefore some protein atoms have to be included in the QM subsystem while others are in the MM subsystem. Then the QM/MM boundary necessarily separates covalently bound atoms and this requires special considerations in setting up the system.

A simple way to separate the system into subsystems is to introduce link atoms [7] into the QM subsystems so as to saturate the dangling bonds cut by the separation (*Figure 11.4*).

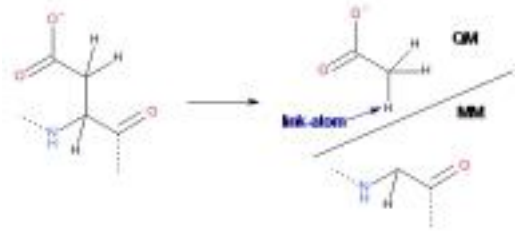


Figure 11.4. Separation of covalently bound subsystems by the introduction of a link-atom

Link atoms are most often H-atoms, but other atoms and chemical groups are also occasionally used. When the cut bond is far enough from the chemical event, then the QM wave function is not expected to be seriously perturbed by the added link-atom. By contrast, the newly introduced link-atom is close to other atoms in the MM system and it may corrupt calculated properties.

Another way of separating covalently bond subsystems is to assign strictly localized molecular orbital (SLMO) to the bond [8],[9],[10] (*Figure 11.5*). An SLMO is formed with 2 hybrid orbitals centred on the bound atoms. Their orbital coefficients are taken from calculations performed for model molecules that include a chemical motif similar to the one in the system investigated. The calculation includes the determination of the wave function for the model molecule, the localization of the orbitals, and the omission of those coefficients of the localized orbital that are not centred on the bound atoms. The resulted strictly localized orbital is renormalized and is used as a frozen orbital, i.e. its coefficients are not optimized in the QM/MM calculation.

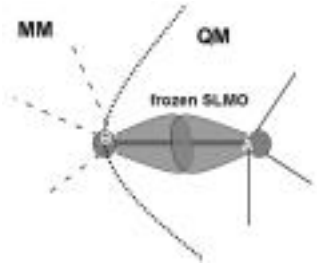


Figure 11.5.Separation of covalently bound subsystems by using strictly localized molecular orbitals

The appropriate selection of the boundary4 between covalently bond subsystems is essential for a sensible application of mixed methods. Either the link-atom or the frozen localized orbital method is used, the system is advantageously cut along a localized non-polar bond, like the C_α-C_β bond in amino acids.

5.3. QM/MM applications

As an illustrative example of QM/MM applications the calculation of the energy curve for the proton transfer between amino acids Asp and His is presented [11]. Note that this calculation does not show the full power of QM/MM approaches. On the other hand, the simplicity of the model allows a clear understanding of the principal features of QM/MM calculations. The system comprises an Asp and a His residue and energy is evaluated as a function of the AspO-H distance as the proton moves from the Asp side chain towards the imidazole of the His (see *Figure 11.6*).

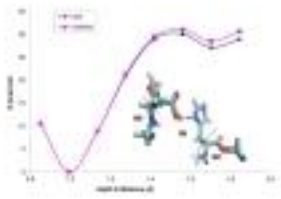


Figure 11.6.Energy of the Asp-His system as a function of the separation of the proton from the O-atom of Asp. System separation into QM and MM regions is also indicated. (Reproduced from ref. [11] with permission.)

In video 11.7 the proton moving is presented between the systems. "Ball-and-Stick" representation shows the QM region while pure stick representation refers to the MM part.

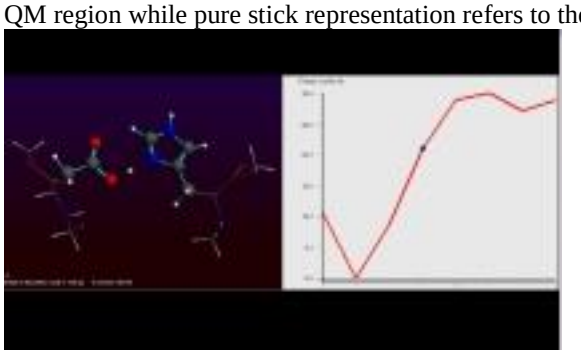


Figure 11.7. QM/MM calculation of proton moving. The "ball-and-stick" representation refers to the QM region while the pure stick part to the MM one.

The QM subsystem includes atoms near the moving proton. The boundaries between the QM and MM subsystems were chosen at the C_β-atoms of both amino acids (C_α and C_β correspond to atoms A and B, respectively, in *Figure 11.5*). The MM atoms in this simplified model are represented by point charges. The C_α–C_β bonds are SLMOs (nonzero orbital coefficients are only on C_α and on C_β atoms). These orbitals are not optimized; rather they are taken from a model calculation performed for a molecule with nonpolar C-C bond. The wave-function for the QM subsystem is evaluated at various AspO-H separations. The energies are shown in *Figure 11.6*. Energies obtained by standard QM calculations for the whole system are also shown in *Figure 11.6* for reference. The QM/MM and reference curves are vertically shifted so that their minimum energy points are superimposed. The shape of the QM/MM curve well follows that of the reference and the positions of the two minima and the maximum in between agree. On the other hand, the relative energy of the maximum and the second minimum is slightly shifted to lower values in the QM/MM curve. In summary, this example illustrates that a QM/MM calculation is able to well reproduce the full QM results at a reduced computational cost. Interested readers are referred to the “Further Readings” section for retrieving several examples and references for QM/MM applications.

6. References

1. C. Cohen-Tannoudji, B. Diu, F. Laloe, Quantum mechanics , Wiley, New York

2. P.R.Halmos Finite-Dimensional Vector Spaces, Princeton University Press, Princeton A. Nagy “Density functional. Theory and application to atoms and molecules” Phys. Rep. 298, 1-79, (1998)

3. A. Nagy “Density functional. Theory and application to atoms and molecules” Phys. Rep. 298, 1-79, (1998)

4. P. Hohenberg, W. Kohn, “Inhomogeneous electron gas” Phys. Rev. B 136, 864-871 (1964)

5. P. A. M. Dirac, "Note on exchange phenomena in the Thomas-Fermi atom". Proc. Cambridge Phil. Roy. Soc. 26, 376–385, (1930)

6. W. Kohn, L. J. Sham, "Self-consistent equations including exchange and correlation effects". Phys Rev A, 140, 1133–1138, (1965).

7. M. J. Field, P. A. Bash, M. Karplus, “A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations”, J. Comput. Chem. 11, 700–733 (1990).

8. A. Warshel, M. Levitt, “Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme”, J. Mol. Biol. 103, 227–249, (1976)

9. V. Théry, D. Rinaldi, J. L. Rivail, B. Maigret, G. G. Ferenczy, “Quantum mechanical computations on very large molecular systems: The local self-consistent field method”, J. Comput. Chem. 15, 269–282, (1994).

10. J. Gao, P. Amara, C. Alhambra, M. J. Field, “A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations”, J. Phys. Chem. A, 102, 4714–4721, (1998).

11. G.G. Ferenczy, „Calculation of Wave-Functions with Frozen Orbitals in Mixed Quantum Mechanics/Molecular Mechanics Methods. Part I. Application of the Huzinaga Equation.” J. Comput. Chem. 34, 854-861, (2013).

7. Further Readings

1. “Ideas of Quantum Chemistry”, Ed.: L. Piela, Elsevier, ISBN: 978-0-444-52227-6 (2007)

2. H. M. Senn, W. Thiel, “QM/MM Methods for Biomolecular Systems”, Angew. Chem. Int. Ed. , 48, 1198-1229, (2009)

3. R. A. Mata, “Application of high level wavefunction methods in quantum mechanics/molecular mechanics hybrid schemes”, Phys. Chem. Chem. Phys., 12, 5041-5052, (2010)

4. S. C. L. Kamerlin, S. Vicatos, A. Dryga, A. Warshel, “Coarse-Grained (Multiscale) Simulations in Studies of Biophysical and Chemical Systems”, Annu. Rev. Phys. Chem., 62, 41-64, (2011)

8. Questions

1. What is theoretical background of the existence of potential energy surfaces?

2. What kind of systems can be calculated by the HF method ?

3. What is the SCF method?

4. Why is the choice of the exchange-correlation potential so important in DFT calculations?

5. Can you imagine such a situation, when the Kohn-Sham orbitals would be equal to the Hartree-Fock ones?

6. Please analyse the PES in *Figure 11.1*. !

7. What type of energy evaluation schemes in QM/MM systems have been proposed?

8. How covalently bound systems are separated into QM and MM subsystems in mixed methods?

9. Glossary

- *Vector space or linear space*: A mathematical structure where a set endowed with two operations called addition and scalar-multiplication which obey certain rules. The elements of the set are called generally vectors. The addition acts between two vectors while scalar multiplication refers to the multiplication of a vector with a real or complex number.

- *Operator (linear operator)*: A mathematical map between two (not necessarily different) vector-spaces with special properties.

- *Eigenvalue-equation, eigenvectors, eigenvalues*: Let an operator which maps over vector-space V. Those vectors (vi) are referred to as eigenvectors (or sometimes eigenfunctions) of operator which are the solutions of the eigenvalue equation . Here λi denotes the corresponding eigenvalue of eigenvector vi .

- *Separation of variables*: A mathematical method of solving many variables differential equations.

- *Born-Oppenheimer approximation*: An important step to decrease the number of variables during the solution of the Schrödinger equation, it fully decouples the motion of nuclei from the electronic system.

- *Potential Energy Surface (PES)*: The PES of a molecule is a hypersurface in a p+1 dimensional coordinate space, where p is the number of geometric parameters which characterize the geometry of the nuclei and the +1 dimension comes from the total energy of the electron system.

- *Boson*: particle with integer spin

- *Fermion*: particle with half spin

- *One-particle (two-particle) operator*: An operator built up from the sum of such operators which depend on the variables of only one (two) particle(s).

- *Antisymmetric wavefunction*: State of a many particle system which changes its sign for the exchange of any two particles.

- *One-particle function*: A function which depends on the variables of only one particle.

- *Functional*: A special operator which assign (real or complex value) scalars to the elements of a vector space. Therefore, it is a map between an arbitrary vector space and the vector space of real (or complex) numbers.

- *Self Consistent Field method (SCF-method)*: An iterative method for solving the one-particle eigenvalue (Hartree-Fock or the Kohn-Sham) equations when the operator in the eigenvalue equation contains implicitly the eigenvectors of the same equation. In the first step an initial guess is applied for the operator. The next step is made by building up the operator from the eigenvectors of the first step and solving again the equation. The iterations should continue until the change in a chosen descriptor (e.g. the eigenvalue of the equation) reaches a certain threshold value.

- *Hartree-Fock method*: A method to solve the Schrödinger equation of the N-electron system. It is based on the solution of a one-particle eigenvalue equation and applies the obtained one-particle eigenfunctions for building the N-electron wavefunction in its Slater-determinant form.

- *Density Functional Theory (DFT)*: Another method to solve the Schrödinger equation of the N-electron system. In DFT the basic quantity is the total density of the electron system instead of the wavefunction. Solving the one-particle Kohn-Sham equations the total density can be determined and from that we can obtain the total energy of the system as well.

- *Mixed methods*: In mixed methods a large system is divided into two (or sometimes three) subsystems and calculates them at different level of theory. Typically, one of the subsystems (the smaller one) calculated at high level of theory and the other at molecular mechanical level. Therefore it is often mentioned as QM/MM method.

- *Link atom*: When subsystems are connected covalently in mixed methods, link atom is introduced to saturate the dangling bond. In general H-atom is applied but other atom and chemical group are also occasionally considered.

- *Strictly Localized Molecular Orbital (SLMO)*: Application of SLMO provides another possibility to separate covalently bonded subsystems. An SLMO is formed by 2 hybrid orbitals centered on the bound atoms. The coefficients of the combination are determined by previous model calculations.

Chapter 12. Evaluation of Reaction Kinetics Data

(Eufrozina A. Hoffmann)

Keywords: reaction rate, reaction mechanism, Michaelis-Menten mechanism, type of inhibitions, Arrhenius equation, kinetic parameter estimation, EC_{50}, IC_{50}

What is described here? This chapter introduces the laws governing rates of reactions, particularly those relevant to biochemical systems. The mechanism of enzyme reactions which plays a key role in biochemistry is discussed in detail. The modern methods of parameter estimation are highlighted giving example to determine EC_{50} and IC_{50}.

What is it used for? To calculate the parameters governing biochemical reactions, to predict temporal behavior, temperature dependency and type of the reacting systems.

What is needed? The knowledge of how to solve basic differential equations. Elementary physical chemistry is also a prerequisite.

1. Introduction

The eventual goal of many computational chemistry projects is contribution to predicting temporal evolution of reactions, *ie*. to modeling their kinetics. Biochemical reactions are governed by the same kinetic laws as simple chemical reactions. There is a difference in their complexity, however. The description of systems typically leads to systems of ordinary differential equations, ODEs. (Recall from mathematics that an ODE is an equation involving a sole independent variable and its derivative(s), but no partial derivatives with multiple variables.) Sometimes, with application of simplifying conditions, analytical solutions can be found. More often a numerical solution is needed - determining of which is a standard computational problem. Some typical cases are covered herein.

The practical importance of studying kinetics laws is twofold. First, they provide a simple theoretical framework within which the behavior of complicated (bio-)chemical systems can be understood. Second, the mathematical models derived from them allow researchers to make predictions; that is,reaction rates can be calculated for as yet unexplored conditions. It is crucial to keep in mind that any such prediction could only be as good as the parameters it is based on: watch out for "garbage in – garbage out" situations. For this reason pitfalls of parameter determination will be elaborated in this chapter, too. The basic treatise presented will not dwelve on the minutia of computational methods for the underlying parameters such as activation energies. Rather, a general overview intended to provide a frame of reference is given.

.

2. Isothermal rate constants

In this chapter the types of the fundamental kinetic equations are summarized. It is customary to define kinetic parameters under constant temperature first, and later deal with their temperature dependence separately.

General differential rate equation

The change with time of a chemical species due to a reaction (termed rate of consumption or appearance, *resp.*, for reactants or products) is expressed mathematically as the derivative with respect to time; in a system of constant volume [1]:

$$r_A = \mp \frac{d[A]}{dt} \tag{12.1}$$

where *a* the square brackets denote concentration of the substance. The sign is negative for reactants and positive for products. The rate of the reaction is defined as

$$r=-\frac{1}{a}r_A \tag{12.2}$$

where *a* is the stoichiometric coefficient for substance A in the reaction in question. The stoichiometric coefficient has negative sign for reactants and positive for products, by convention; therefore the rate is always non-negative.

The so-called partial order of reaction, in respect of species A, can mathematically defined as

$$n_A=\mp\frac{d\ln r}{d\ln[A]} \tag{12.3}$$

Eq. 12.3 defines the quantity regardless of the rate law. In complicated cases this "apparent" order may be a function of the progress of reaction, as well as of the concentration(s). Under these circumstances it is not useful to speak of order of reaction, according to the IUPAC recommendation [1]. Traditionally the definition had been tied to the generic rate law of the form $r=k\prod_i[A_i]^{\alpha_i}$; here the partial orders are simply the exponents α_i. The overall order of reaction is the sum of all partial orders.

Special integrated rate equations

Many systems of practical importance have simple analytical (closed-form) solutions. These are discussed in the following sub-sections.

First-order reaction

The simplest case is the first-order reaction:

$$-\frac{d[A]}{dt}=k[A] \tag{12.4}$$

Direct integration yields the solution of this differential equation as:

$$\ln[A]=\ln[A]_0-kt \tag{12.5}$$

which can be transformed into:

$$[A]=[A]_0e^{-kt} \tag{12.6}$$

Pseudo-first order reaction

An often utilized experimental technique for studying non-unimolecular reactions is to make them pseudo-first order [2]. This means making all but one concentration constant (usually with keeping the others in large excess). If the corresponding partial order is one, then the rate equation becomes formally first-order.

$$r=k[A_1]\prod_{i>1}[A_i]^{n_i}\simeq k_1[A_1] \tag{12.7}$$

Just like in the case of true first order reaction, this differential equation can be easily integrated to:

$$\ln[A_1]=-k_1t \tag{12.8}$$

The simplicity of evaluating results made this the preferred method of studying mechanisms in many cases. A great advantage of the pseudo-first order case (just like that of true first order) is that only relative concentration is needed for determining the rate constant. Conversely, from a given rate constant the percentage yield at any time can be calculated without knowing the initial concentration.

Higher-order reactions

Simple bimolecular reactions have second-order rate law. In case of a single reactant the differential form is: $r= k[A]^2$, from which the following integrated equation can be derived: $1/[A]_t- 1/[A]_0= kt$. When two different reactants, each with partial order of one, occur then the differential rate equation is: $r= k[A][B]$. Integrating this yields different forms depending on whether the initial concentrations are equal. If $r= k[A]_0[B]_0$, then the two concentrations remain equal all the time, and $1/[A] - 1/[A]_0 = 1/[B] - 1/[B]_0 = kt$ (note that this is the same as in

the previous case, due to the equivalence of the two reactants). If the two initial concentrations differ from each other, then the following integrated equation is obtained: $1 / ([A]_0 - 1/[B]_0) \ln([A][B]_0)/([B][A]_0) = kt$; this can be rearranged to: $\ln [A]/[B] = \ln [A]_0/[B]_0 + k ([A]_0 - [B]_0)t$

.

Zero-order reactions

For the sake of completeness zero order should be mentioned. This occurs when the rate is limited by some factor other than reactant concentration. Examples include photochemical reactions governed by the number of photons, or surface reactions at near-full coverage. Biochemically relevant systems include proton-catalyzed reaction whose rate is determined by the hydrogen ion concentration which is not encountered in the rate equation. When the controlling factor (pH in the latter case) is constant then the rate does not change regardless of varying the concentration of the reactant(s).

Coupled multi-reaction systems of biochemical interest

Michaelis-Menten mechanism

There are many systems of interest that contain multiple different reactions, coupled *via* sharing common species. One of the most widely known systems with this property is the Michaelis-Menten mechanism of enzyme-substrate interactions, introduced in 1913 (*Figure 12.1*) [3]. According to this mechanism the enzyme (E) and the substrate (S) forms a complex (C) in a pre-equilibrium process, after which this complex converts to a product (P) that does not bind to the enzyme.

Figure 12.1. Michaelis-Menten mechanism

Considering the time dependence of the various species, four differential equations can be written:

$$\begin{aligned}
\frac{d[S]}{dt} &= k_{-1}[C] - k_1[E][S] \\
\frac{d[E]}{dt} &= (k_{-1}+k_2)[C] - k_1[E][S] \\
\frac{d[C]}{dt} &= k_1[E][S] - (k_{-1}+k_2)[C] \\
\frac{d[P]}{dt} &= k_2[C]
\end{aligned} \qquad (12.9)$$

This system of differential equations can be solved numerically for specific initial concentrations, or analytically applying certain initial conditions and assumptions. The generally applied initial conditions are that the initial concentrations $[E]_0$, $[S]_0$, and $[C]_0=0$ are given $[E]=[E]_0-[C]$, yielding two independent differential equations.:

$$\begin{aligned}
\frac{d[S]}{dt} &= k_{-1}[C] - k_1([E]_0 - [C])[S] \\
\frac{d[C]}{dt} &= k_1([E]_0 - [C])[S] - (k_{-1}+k_2)[C]
\end{aligned} \qquad (12.10)$$

To solve these equations Michaelis and Menten applied the rapid equilibrium approximation that after a short time: $d[S]/dt \simeq 0$ thus

$$[C] = \frac{[E]_0[S]}{K_S+[S]}, \text{ where } K_S = \frac{k_{-1}}{k_1} \text{ and } \frac{d[P]}{dt} = k_2\frac{[E]_0[S]}{K_S+[S]} \qquad (12.11)$$

A more commonly applied solution of the Eqs. 12.10, called Quasi Steady State Approximation (QSSA), was developed by Briggs and Haldane [4]. They assumed that the concentration of the substrate-bound enzyme changes much more slowly than those of the product and substrate. Therefore $d[C]/dt \simeq 0$ Applying this assumption:

$$[C] = \frac{[E]_0[S]}{K_M+[S]} \text{ and } \frac{d[P]}{dt} = v_{max}\frac{[S]}{K_M+[S]}, \text{ where } K_M = \frac{k_{-1}+k_2}{k_1} \text{ and } v_{max} = k_2[E]_0 \qquad (12.12)$$

K_M is called Michaelis constant and v_{max} is the maximum rate. These are important parameters to characterize enzyme inhibitions. The rate of the product formation is: $v = k_{cat}[C]$. In the Michaelis-Menten mechanism, where there is only one enzyme-substrate complex and all binding step are fast, k_{cat} is the first rate constant for the chemical conversion of the enzyme-substrate complex to the enzyme-product complex [5]. In the QSSA, when the dissociation of C is fast: $k_{cat}=k_2$; but when it is far slower than the rate of the chemical steps k_{cat}=constant. In the case of extended mechanism of Michaelis-Menten scheme, where additional intermediates occur, K_M and k_{cat} are combinations of various rate and equilibrium constants [5].

The importance of k_{cat} that it represents the maximum number of substrate molecules converted to products per active site per unit time. It is often referred as "turnover number". k_{cat}/K_M is called as "specificity constant", it determines to the properties and the reactions of free enzymes and free substrate.

In the past, several traditional graphical evaluations of Michaelis-Menten mechanism were popular, such as Hanes-Woolf plot [6], Lineweaver-Burk plot [7], Eisenthal-Bowden plots [8], in order to obtain kinetic parameters; these methods may cause statistical bias due to the transformations applied, and their use is deprecated by the availability of direct numerical parameter estimation with computers. One of the first computerized evaluations was the work of Sakoda et al [9]. They obtained the best-fit values of the K_m and v_{max} in the Michaelis-Menten equation by the method of least squares with the Taylor expansion for the sum of squares of the absolute residual. Raaijmakers applied the method of maximum likelihood for analysis of enzyme kinetic experiments which obeys Michaelis Menten mechanism [10]. The strong boundary assumptions in the QSSA itself have been modified by some authors, achieving better agreement with experiments this way.

Borghans *et al.* [11] developed the so called total QSSA (tQSSA) method. Their proposition was that, for conditions when the total enzyme concentration ($[E]_T$) and the initial substrate concentration are comparable, the proper intermediate timescale variable is $[\hat{S}(t)] = [S(t)] + [C(t)]$. In terms of this variable, the governing equations are:

$$\begin{aligned} &[E]+[C]=[E_T]=\text{constant} \\ &[E_T][\hat{S}]-([E_T]+[\hat{S}]+K_m)[C]+[C]^2=0 \\ &\frac{d[\hat{S}]}{dt}=-k_2[C] \end{aligned} \qquad (12.13)$$

A practical method of analysis of Michaelis-Menten mechanism was developed by Garneau-Tsodikova *et al.* [12]. Their formalism does not involve any other approximations such as the steady-state, limitations on the reactant concentrations or on reaction times. Based on the total concentration of the enzyme and on the total concentration of the substrate, they derived the concentration of the enzyme-substrate complex. This was substituted into the kinetic rate equation of product formation. A differential expression so obtained can be integrated to yield the general solution in a closed analytical form.

So far we have discussed the Michaelis-Menten mechanism in detail. There are enzymes whose kinetics can only be described properly by some more complicated mechanism. One such case is when multiple substrates can bind, which occurs frequently in nature.

The *Ordered Sequential Mechanism* (*Figure 12.2.a*) is very similar to the Michaelis-Menten scheme. However, binding of the second substrate is subsequent to the binding of the first substrate, in a separate equilibrium. The molecular explanation to this is that conformation change, induced by the binding of the first substrate, makes possible the binding of the second substrate. The sequence of the product formation is also determined. A special case of the Ordered Mechanism is the *Theorell-Chance Mechanism* (*Figure 12.2.b*), in which ternary complex does not accumulate, and two products are formed. These products are different type of molecules. In the *Random Sequential Mechanism* (*Figure 12.2.c*) either binding site can bind the first substrate, and the remaining free site then binds the second substrate. The ternary complex so formed is releases the product while freeing the enzyme. In the *Ping-Pong* (or *Substituted Enzyme*, or *Double-Displacement*) *Mechanism* (*Figure 12.2.d*), the reaction of the first substrate with the enzyme covalently modifies the enzyme, and one product is formed. This modified enzyme reacts with the second substrate yielding the second product.

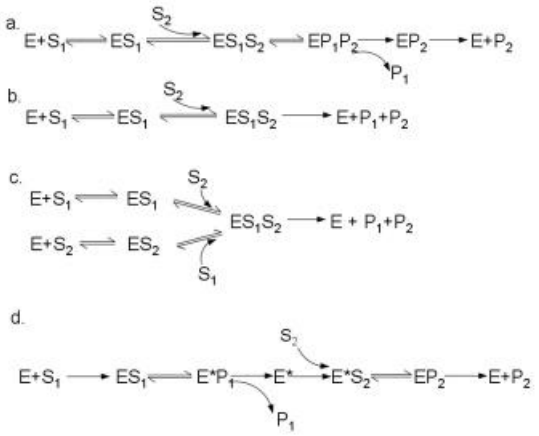


Figure 12.2. Enzyme mechanisms which do not follow Michaelis-Menten scheme (a, Ordered Sequential Mechanism; b, Theorell-Chance mechanism; c. Random Sequential Mechanism; d, Ping-Pong Mechanism)

Several recent experiments indicate that the behavior of many enzymes is more complicated. In these cases, computer simulation based on experimental data can help to build the kinetic mechanism.

Types of inhibition

In the former section the kinetic of simple enzyme reactions has been explained. There are molecules which bind to an enzyme decreasing its activity. They are called inhibitors. Similarly to enzyme-substrate inhibition, the receptor-ligand kinetics can also be inhibited, and the same terms also used in this respect.

Several well known drug molecules are enzyme inhibitors. For example methotrexate [13], an inhibitor of dihydrofolate reductase, is frequently applied in cancer chemotherapy and in autoimmune diseases.

The inhibitions can be either reversible or irreversible according to the type of binding.

Reversible inhibitors do not react chemically with the enzyme, and in most cases they can be easily removed by dilution or dialysis. This is because inhibitors bind to enzymes with weak bonds. Several of these bonds together form a strong and specific binding, however. Reversible inhibitors can be further classified [14]:

- *Competitive inhibitors* compete with the substrate for the active site of the enzyme. Generally, they have similar structure to the real substrate; and if the concentration of the substrate is large enough, the competitive inhibition can be overcome. They do not bind to the enzyme-substrate complex already formed.The binding efficiency (K_m) is increased in case of competitive inhibition because the inhibitor interferes with substrate binding); but catalysis in ES is not slowed because the inhibitor cannot bind to the complex, therefore maximum velocity (V_{max}) is not affected.

- *Uncompetitive inhibitors* bind only to the substrate-enzyme complex, but do not interact with the free enzyme molecules, thus the inhibition cannot be reduced by increasing concentrations of substrate. Therefore both V_{max} and K_m decrease.

- *Mixed inhibitors* can bind both to the enzyme and to the enzyme-substrate complex as well. Therefore this type of inhibition can be reduced, but not overcome by the large amount of substrate. Sometimes mixed inhibition is due to an allosteric effect (*allosteric inhibition*), where the inhibitor binds to a different - allosteric- site on an enzyme and this leads to an altered conformation of the enzyme where the substrate no longer fits. Another type of mixed inhibitiors is *non-competitve inhibitors.* Their binding to the enzyme reduces its activity without altering the affinity of the enzyme toward the substrate. Therefore, the extent of inhibition is solely determined by the concentration of the inhibitor.V_{max} is lowered, but K_m is not changed.

Irreversible inhibitors usually bind with covalent bond to the enzyme. This type of inhibition cannot be reversed, and it follows neither competitive nor non-competitive kinetics. Sometimes it is difficult to decide whether an inhibition is irreversible, or reversible with tight binding of the inhibitor making the enzyme released very slowly. This latter type of inhibitors is called tight-binding inhibitors. If an enzyme has two or more active sites, the inhibitors can show different type of kinetic: for example, competitive inhibition on one binding site and non-competitive on another one [15].

3. Temperature dependence of rateconstant

Arrhenius equation

Changing the temperature dramatically affects reaction between either ligands and receptors, or enzymes and substrates. On the one hand, all the equilibrium and rate constants in the mechanisms are temperature dependent. On the other hand, the structure of biomolecules may change with temperature, thus their binding and activity can get modified.

The temperature dependence of the equilibrium constant (K) is described by the van't Hoff equation [16]:

$$\frac{d\ln K}{dT}=\frac{\Delta H^{\Theta}}{RT^2} \qquad (12.14)$$

where T denotes the temperature, R is the gas constant, ΔH^{θ} indicates the standard enthalpy change for the process.

Over temperature intervals where the reaction enthalpy can be considered constant, the above equation can be integrated to yield:

$$\ln\left(\frac{K_2}{K_1}\right)=\frac{\Delta H^{\theta}}{R}\left(\frac{1}{T_1}-\frac{1}{T_2}\right) \qquad (12.15)$$

The empirical formula for the temperature dependence of rate constants was also suggested by van't Hoff (in 1884 [17]), and was given a physical interpretation based on the collisional theory of gases by Arrhenius (in 1889 [18]). According to the equation named after him:

$$k=Ae^{\frac{-E_A}{RT}} \qquad (12.16)$$

where A is the pre-exponential factor, E_A is the so-called activation energy. In a simplified picture, A gives the total number of collisions and the $\exp(-E_A/RT)$ is the probability that any given collision will result in a reaction.

Extended Arrhenius formulas

The Arrhenius equation was modified by several authors [19]. The most important theoretical treatments of reaction rates are the so-called t(TST) − also known as activated complex theory − introduced by Eyring, Evans and Polanyi [20], and by Pelzer and Wigner [21] in the 1930s.

The principal result of TST is the formula for the rate constant k:

$$\ln\left(k\frac{\hbar}{k_B T}\right)=-\Delta\frac{G^{\ddagger}}{RT} \qquad (12.17)$$

where $\Delta G^{\ddagger}$ is the free enthalpy of activation (describing the activation complex characteristic for the reaction), k_B, ħ and R are the Boltzmann, Planck and gas constants, respectively.

The traditional phenomenological handling of the Arrhenius equation plot, $\ln(k)$ *vs.* $1/T$. The slope of this, usually (nearly) linear plot is the empirical activation energy E_A. Applying the procedure to the TST formula shows that the slope is:

$$\frac{-E_A}{R}=\frac{\partial\ln k}{\partial(1/T)}=T-\frac{1}{RT}+\frac{\partial\ln H^{\ddagger}}{\partial(1/T)}+\frac{\Delta H^{\ddagger}}{R}+\frac{1}{R}\frac{\partial\Delta S^{\ddagger}}{\partial(1/T)} \qquad (12.18)$$

The temperature dependence of $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ is usually negligible over the range of T measured, so the above expression simplifies to the formula:

$$\frac{-E_A}{R}=\frac{\partial \ln k}{\partial(1/T)}=T-\frac{\Delta H^{\ddagger}}{R} \quad (12.19)$$

The use of the Arrhenius equation is widespread in the literature. In fact, if the overall rate constant follows the Arrhenius equation, then this is occasionally considered evidence that the structure of biomolecules (receptor or enzyme) remained unchanged throughout the temperature range investigated. This reasoning is faulty, however: without a detailed knowledge of the mechanism, the appearance of a single Arrhenius equation is insufficient to exclude variations of the structure in question. There are also cases when a complex mechanism cannot be described even with extended Arrhenius-type formulas. Although detailed discussion of these cases is beyond the scope, it is important to keep in mind that this is typical rather than exceptional behavior when more than a single reaction dominates. It should also be noted that state-of-the-art computational methods, such as high-level molecular dynamics simulations, have matured to the point where accurate theoretical predictions can be made for the transition states and their associated activation energies for some enzyme catalyzed reactions - see, e.g., [22] for a recent example.

4. General remarks on parameter estimation

Traditionally kinetic models used to be linearized whenever possible, in order to estimate their parameters. Severe shortcomings of this approach have long been recognized [23]. It is important to calculate the statistical uncertainty of the parameters obtained, based on the experimental errors inherent in the input data. This becomes complicated when linearization is involved, and the extra effort needed for proper calculation negates the apparent simplicity of using linear regression. With advanced nonlinear methods readily available today these transformation are unnecessary. Direct parameter estimation is preferred. A recent theoretical paper by Tasi and Barna [24] elaborates this on the example of a Michaelis-Menten mechanisms evaluated according to the Woolf-Lineweaver-Burk form. They showed that non-linear least-squares fitting can be adequately handled with their method, which is based on the simplex optimization technique with error estimation.

There is a large variety of software available. For relatively simple systems, even using built-in nonlinear solvers for spreadsheet programs is feasible. General-purpose mathematical suites, such as Mathematica, MATLAB or their freeware equivalents, can also be used. For example, there are several specialized kinetics packages available in the “R” free software environment - for a current listing browse

http://cran.r-project.org/web/packages/available_packages_by_name.html.

Turányi *et al.* [25] reported developing in-house MATLAB code for encoding sophisticated statistical evaluation of large-scale kinetics models. There is also a number of special-purpose programs aimed at kinetics. One of the most comprehensive such suites is ZiTa [26]. This incorporates ODE-solver capability with parameter estimation, and utilizes flexible model definition that accommodates equilibrium equations besides kinetic ones.

We have seen that even a relatively simple mechanism such as Michaelis-Menten leads to differential equations whose solution can be different, depending on the boundary conditions and on the approximations applied. The majority of enzyme reactions does not exactly follow Michaelis-Menten kinetics [27], and many reaction systems are more complicated. These types of problems can only be treated via *simulation*: with numerical solution of the system of differential equations written according to an assumed mechanism, and comparing the modeled results with experiments. Numerically solving the system of ordinary differential equations (ODEs) is a standard computational problem [28]. A particular problem that frequently occurs in kinetics, when there are unstable intermediates or other fast reacting species, is the so-called stiffness: the solution involves terms that may vary exceedingly rapidly, along with slower steps. This cause numerical instabilities, poor convergence, and undue restrictions on the step size applicable. There are many well-tested algorithms [29] and program implementations [30] available for overcoming these difficulties and routinely solving stiff ODEs in kinetic systems, however. While historically these tools were mostly developed by, and distributed to, users focusing on gas-phase kinetic applications, in recent years there have been growing awareness for their utility in the fields of biochemistry and systems biology, as well. It is now well recognized in enzyme kinetics, for example, that the co-existence of fast processes (like enzyme-substrate interaction) with slow steps (such as typical product formation) causes stiffness of the mechanism. There are now programs specifically targeted for biochemical research – and many of these tools are free for academic purposes. For a sampling of their continuously expanding range see the references [30e], and citations therein.

5. Parameter estimation in pharmacokinetics

The scientific discipline investigating rates of processes in pharmacology is called pharmacokinetics. Much of the mathematical formalism developed for chemical kinetics finds application here, even though biological transformations as well as pure physical processes also play important roles besides chemistry. Many effects of interest, typically depicted as dose-response curves, follow sigmoidal shape. One frequently used form is the Hill equation [32]:

$$\theta = \frac{[L]^{n_H}}{K_D + [L]^{n_H}} \tag{12.20}$$

where θ is the fraction of maximum (either for bound ligand or response). Application of Eq. 12.20 marked the start of quantitative treatment in pharmacology [33].

It is often desirable to describe these curves with a single parameter, for which the mid-point is the most frequently used practical choice. For active compounds (*agonists*), the term EC_{50} is defined as the concentration that produces 50% of the maximal possible effect. For inhibitors, the 50% inhibition concentration, IC_{50}, is used: that is the concentration of the inhibitor which reduces an effect (such as some response, or binding of the agonist) to half its maximum value. When the curve is symmetrical then the half-maximum level coincides with the inflection point. There are cases, however, when asymmetrical curves are encountered where the inflection point is distinct from the mid-point, so care should be taken not to confuse the two.

Note that Eq. 12.20 assumes a zero baseline. Switching to a generic dependent variable y instead of θ, allowing for a non-specific y_0 effect present at $[L]=0$, and rearranging for a traditional linearized plot yields

$$\ln\left(\frac{y_M - y_0}{y - y_0} - 1\right) = n_H \ln(K_S) + n_H \ln([L]) \tag{12.21}$$

(where $K_S = K_D^{n}$ was introduced, and y_M is the top asymptote). This equation (or one of its equivalent forms) is, mathematically, a four-parameter logistic curve. n_H is called Hill slope (or coefficient). The mid-point is at $[L]=K_S$.

A simulated dataset is used to illustrate fitting to Eqs. 12.20-21. Plotted below, the theoretical curve is characterized by IC_{50}=10.0 μM, n_H=1.00, $K_D=K_S$=10.0 μM. The data points shown are generated with 10% relative error. Non-linear parameter estimation based on Eq. 12.20 yields n_H=1.00 and K_D=9.93 μM (dashed curve), i.e. IC_{50}=9.93 μM. Linearized fitting *via* Eq. 12.21 yields n_H=1.01 and K_S=9.72 μM (dotted curve), *i.e.* IC_{50}=9.72 μM.

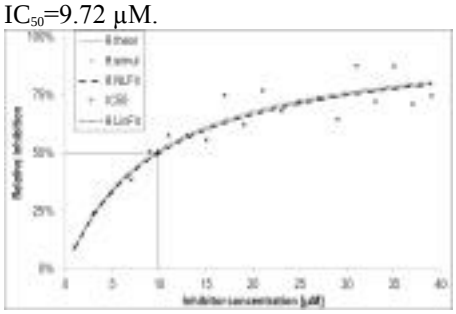


Figure 12.3. A simulated dataset is used to illustrate fitting to Eqs. 12.20-21

Due to the importance of the EC_{50} (or IC_{50}) parameter in pharmaceutical research, many specialized software tools are available for its determination. Just as mentioned in the section on kinetic parameter estimation, it is crucial to use proper statistical treatment of the experimental errors, and for this reason direct non-linear algorithms should be preferred over deprecated linearization methods [34].

6. References

1. IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook

2. a) M. C. Sauer Jr and B. Ward. Reactions of hydrogen atoms with benzene and toluene studied by pulsed radiolysis: reaction rate constants and transient spectra in the gas phase and aqueous solution. J. Phys. Chem. 71, 3971-3983 (1967).

   b) V. D. Parker, W. Hao, Z. Li and R. Scow, Nonconventional versus conventional application of pseudo- first- order kinetics to fundamental organic reactions. Internat. J. Chem. Kinet., 44, 2-12 (2011).

3. L. Michaelis and M. L. Menten, Die Kinetik der Invertinwirkung. Biochem. Z. 49, 333-369 (1913).

4. G. E. Briggs, J. B. S. Haldane, A note on the kinetics of enzyme action. Biochem. J. 18, 338-339 (1925).

5. A. Fersht, Structure and Mechanism in protein science, W. H. Freeman and Company, New York (2000).

6. C. S. Hanes, Studies on plant amylases: The effect of starch concentration upon the velocity of hydrolysis by the amylase of germinated barley. Biochem. J. 26, 1406 (1932).

7. H. Lineweaver and D. Burk, The Determination of Enzyme Dissociation Constants. J. Am. Chem. Soc. 56, 658-666 (1934).

8. R. Eisenthal and A. Cornish-Bowden, The direct linear plot. A new graphical procedure for estimating enzyme kinetic parameters. Biochem. J. 139, 715-720 (1974).

9. M. Sakoda and K. Hiromi, Determination of the best-fit values of kinetic parameters of the Michaelis-Menten equation by the method of least squares with the Taylor expansion. J. Biochem. 80, 547-555 (1976).

10. J.G.W. Raaijmakers, Statistical analysis of the Michaelis-Menten equation. *Biometrics*, *43*, 793-803 (1987).

11. J. A. Borghans, R. J. de Boer, L. A. Segel, Extending the quasi-steady state approximation by changing variables. Bull. Math. Biol. 58, 43-63 (1996).

12. S. Garneau-Tsodikova, I. A. Shkel, O. V. Tsodikov, Exact and user-friendly kinetic analysis of the two-step rapid equilibrium Michaelis-Menten mechanism. Anal. Biochem. 387, 276-279 (2009).

13. J. J. McGuire, Anticancer antifolates: current status and future directions. Curr. Pharm. Des. 9, 2593-613 (2003).

14. S. E. Szedlacsek and R. G. Duggleby, Kinetics of slow and tight-binding inhibitors. Meth. Enzymol. 249, 144-180 (1995).

15. I. H. Segel, Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. Wiley-Interscience; New edition (1993).

16. P. W. Atkins, Physical Chemistry 6th edition, Oxford University Press.

17. J. H. van't Hoff, Études de dynamique chimique. F. Muller & Co., Amsterdam (1884). Studies of Chemical Dynamics, revised edition with additions by E. Cohen, translated by T. Ewan (1896); reprinted in 2010 by BiblioBazaar, LLC.

18. S. Arrhenius, Über die Reaktionsgeschwindigkeit bei der Inversion von Rohzucker durch Säuren. Zeits. Phys. Chem. 4, 226-248 (1889). On the reaction velocity of the inversion of cane sugar by acids, translated and published in Margaret H. Back & Keith J. Laidler (Eds.): Selected Readings in Chemical Kinetics, Pergamon Press, Oxford (1967).

19. J. N. Murrell, Understanding the Rates of Chemical Reactions, in: Fundamental World of Quantum Chemistry, Vol. 2, E. J. Brändas, E. S. Kryachko (Eds.), Kluwer Academic Publishers, p155-180 (2003).

20. a) H. Eyring and M. Polanyi, Über einfache gasreaktionen (On Simple Gas Reactions). Z. Physik. Chemie, B12, 279 (1931). Part of this is translated in: M. H. Back, and K. J. Laidler, Selected Readings in Chemical Kinetics, Pergamon Press (1967).

   b) M. G. Evans, M. Polanyi, Some applications of the transition state method to the calculation of reaction velocities, especially in solution. Trans. Faraday Soc. 31, 875 (1935).

21. H. Pelzer and E.Wigner, Über die Geschwindigkeitkonstante von Austausch Reaktionen. Z. Physik. Chem., B15, 445-453 (1932).

22. B. Kormányos, A. K. Horváth, G. Peintler and I. Nagypál. Inherent pitfalls in the simplified evaluation of kinetic curves. J. Phys. Chem. A 111, 8104-8109 (2007).

23. M. W. Lee and M. Meuwly, "Molecular dynamics simulation of nitric oxide in myoglobin." J. Phys. Chem. B 116, 4154-4162 (2012).

24. Gy. Tasi and D. Barna, "Analytical and numerical computation of error propagation of model parameters" J. Math. Chem. 49, 1322-1329 (2011).

25. T. Turányi, T. Nagy, I. Gy. Zsély, M. Cserháti, T. Varga, B. T. Szabó, I. Sedyó, P. T. Kiss, A. Zempléni and H. J. Curran, "Determination of rate parameters based on both direct and indirect measurements" Internat. J. Chem. Kinet. 44, 284-302, (2012).

26. G. Peintler “ZiTa, Version 5.0, a Comprehensive Program Package for Fitting Parameters of Chemical Reaction Mechanism”, University of Szeged, Hungary, 1989-1998. <http://www.staff.u-szeged.hu/~peintler/enindex.htm>.

27. C. M. Hill, R. D. Waightm and W. G. Bardsley, “Does any enzyme follow the Michaelis-Menten equation?”, Molec. Cellular Biochem. 15, 173-178 (1977).

28. G. Dahlquist, "A special stability problem for linear multistep methods", BIT 3, 27-43 (1963).

29. a) B. L. Ehle, “On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems”, Report 2010, University of Waterloo (1969).

b) E. Hairer and G. Wanner, “Solving ordinary differential equations II: Stiff and differential-algebraic problems” (second ed.), Berlin, Springer Verlag, (1996).

c) A. Iserles and S. Nørsett, Order Stars, Chapman and Hall, (1991).

d) G. Wanner, E. Hairer, S. Nørsett, "Order stars and stability theory", BIT 18, 475-489 (1978).

e) S. D. Cohen, A. C. Hindmarsh , “CVODE, a stiff/nonstiff ODE solver in C”, Comput. Phys. 10, 138-143 (1996).

30. a) I. Havlik, J. Votruba, “GASP/S: A GASP IV version with a stiff-ODE integrator”, SIMULATION, 50, 230-238 (1988).

b) R. Macey, G. Oster and T. Zahnley, “Berkeley Madonna User's Guide”, Version 8.0, University of California, Berkeley (2000).

c) M. Okamoto, Y. Morita, D. Tominaga, K. Tanaka, N. Kinoshita, J-I. Ueno, Y. Miura, Y. Maki, and Y. Eguchi, “Design of Virtual-Labo-System for Metabolic Engineering: Development of Biochemical Engineering System Analyzing Tool-kit (BEST KIT)”, Comput. Chem. Engng,, 21(Suppl.), 5745-5750 (1997).

d) F. Perez Pla, J. J. Baeza Baeza, G. Ramis Ramos and J. Palou, “OPKINE, a multipurpose program for kinetics”, J. Comput. Chem., 12, 283-291 (2004).

e) T. Turányi, “Sensitivity analysis of complex kinetic systems. Tools and applications”, J. Math. Chem., 5, 203-248 (1990).

f) A. S. Tomlin, T. Turányi and M. J. Pilling, “Mathematical tools for the construction, investigation and reduction of combustion mechanisms”, Chapter 4 (pp. 293-437) in: Low-temperature Combustion and Autoignition, M.J. Pilling (Editor); Vol. 35 of series 'Comprehensive Chemical Kinetics' Elsevier, Amsterdam, (1997).

31. a) P. Dhar, T. C. Meng, S. Somani, L. Ye, A. Sairam, M. Chitre, Z. Hao and K. Sakharkar, “Cellware - a multi-algorithmic software for computational systems biology”, Bioinformatics, 20, 1319-1321 (2004).

b) D. J. Higham, “Modeling and simulating chemical reactions”, SIAM review, 50, 347-368 (2008).

c) B. Aleman-Meza, Y. Yu, H. B. Schüttler, J. Arnold and T. R. Taha, “KINSOLVER: A simulator for computing large ensembles of biochemical and gene regulatory networks”, Computers and Mathematics with Applications, 57, 420-435 (2009).

d) P. Gonnet, S. Dimopoulos, L. Widmer and J. Stelling, “A specialized ode integrator for efficient computation of parameter sensitivities”, BMC Systems Biology, 6, 46 (2012).

32. a) A. V. Hill, "The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves", J. Physiol. 40 (Suppl), iv-vii (1910).

b) H. Prinz, “Hill coefficients, dose-response curves and allosteric mechanisms. Journal of chemical biology”, 3, 37-44 (2010).

c) S. Goutelle, M. Maurin, F. Rougier, X. Barbaut, L. Bourguignon, M. Ducher and P. Maire, “The Hill equation: a review of its capabilities in pharmacological modeling”, Fundamental and clinical pharmacology, 22, 633-648 (2008).

d) N. Bindslev, “Hill in hell”, Drug-Acceptor Interactions, 257-282 (2008).

33. D. Colquhoun, “The quantitative analysis of drug-receptor interactions: a short history”, Trends Pharmacol. Sci., 27, 149-157 (2006).

34. R. R. Neubig, M. Spedding, T. Kenakin and A. Christopoulos, "International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology", Pharmacol. Rev. December 55, 597-606 (2003).

7. Further Readings

1. 1. G. L. Patrick, An Introduction to Medicinal Chemistry, 3rd Edition, Oxford University Press, 2005.

2. 2. A. G. Marangoni, Enzyme Kinetics: A Modern Approach, John Wiley and Sons., Hoboken, New Jersey, USA 2003.

3. 3. H. M. Sauro, Enzyme Kinetics for Systems Biology, Ambrosius Publishing and Future Skill Software (2011).

8. Questions

1. How do you define the rate of a chemical reaction?

2. Define the partial order of a reaction with respect to species “A” mathematically.

3. What is the difference between a first order and a pseudo-first order reaction?

4. A reaction is known to follow first order kinetics, with rate constant 100 s^{-1}.Calculate the time needed to reach a) 1%; b) 10%; c) 90% conversion.

5. What kind of steps lead to the product formation in an enzyme reaction according to the Michaelis-Menten mechanism?

6. How does the enzyme concentration change with time according to the Michaelis-Menten mechanism?

7. How the concentration of substrate change with time can be expressed in an enzyme catalised reaction if it can be descriebed by the Michaelis-Menten Scheme?

8. An enzyme catalyzed reaction follows Michaelis-Menten mechanism, with K_M=30 mmol/L and V_{max}=5 µmol/s. Calculate the substrate concentration at which the reaction rate is a) 90%; b) 10%; c) 1% of the maximum rate.

9. What is the initial assumption of Quasi Steady State Approximation (QSSA)?

10. Express K_M and vmax based on QSSA model. Describe these two parameters in case of enzyme inhibition.

11. What is the meaning of the chemical expressions “turnover number“ and “specificity constant”?

12. Draw the scheme of the following mechanisms: a, Ordered Sequential; b, Theorell-Chance; c, Random Sequential; d, Ping-pong Mechanism.

13. List the types of reversible inhibition.

14. How do values of V_{max} and K_M change in case of mixed inhibition? Why?

15. What does ’’Allosteric inhibition” mean?

16. What are the main characteristics of irreversible inhibition?

17. How can you express the temperature dependence of the equilibrium constant (K) (van’t Hoff equation)? What is its integrated form if the reaction enthalpy can be considered constant?

18. Describe the linearized form of Arrhenius equation.

19. Predict the percentage change in the rate for a reaction with Arrhenius activation energy 50 kJ/mol.

20. Explain Transition State Theory.

21. How can you express the activation energy based on Transition State Theory?

22. Define EC_{50} and IC_{50}.

23. Explain why the „Hill slope” can alternatively be called „Hill exponent”.

9. Glossary

(Arrhenius) activation energy: an empirical parameter characterizing the exponential temperature dependence of the rate coefficient *k*:

$E_A = RT^2 \frac{\mathrm{d}\ln(k)}{\mathrm{d}T}$, where *R* is the gas constant and *T* the thermodynamic temperature.

Catalyst: a substance that increases the rate of a reaction without modifying the overall standard Gibbs energy change in the reaction.

Elementary reaction: a reaction for which no reaction intermediates have been detected or need to be postulated in order to describe the chemical reaction on a molecular scale.

Enzyme: Bio-macromolecule that functions as ***catalyst*** by increasing the rates of certain biochemical reactions.

Inhibitor: a substance that diminishes the rate of a chemical reaction.

Rate coefficient: the concentration-independent parameter in the rate law of the form $r = k\prod_i [A_i]^{\alpha_i}$.

Rate constant: a ***rate coefficient*** referring to an ***elementary reaction***. Note that it is “constant” only with respect to concentrations, but does depend on external conditions such as temperature or ionic strength.

Chapter 13. Case Studies. Applications to biochemical problems.

(Gábor Paragí, Ferenc Bogár)

Keywords: REMD, protein

What is described here? In this chapter we give a few examples of the application of the methods described in the previous sections. In the case studies we discuss two typical problems.

- In the first one a conformational analysis of three protonated forms of histamine is performed at quantum mechanical level of theory. The conformational spaces are scanned systematically by two dihedral angles and the total energy of the system is plotted as the function of the two parameters. The Potential Energy Surfaces (PES-s) are prepared and the molecular geometry of selected stationary points is presented in pictures.
- The second topic is a molecular dynamical refinement and stability investigation of an experimental polypeptide structure, the Trp-cage miniprotein.

What is it used for? These case studies show the machinery of biomoleculer modelling at work, specifically:

- The investigation of histamine demonstrates how the results of quantum mechanical calculations can be involved into the conformational analysis of a chosen molecule.
- We often use biomolecular structures obtained either from experiments (like X-ray or NMR) or from theoretical predictions (like homology modelling). However, these structures frequently need refinements, because the experimental conditions are far from the physiological ones or the quality of the homology model is not good enough for the further investigations etc. The molecular dynamics is often used for this purpose.

What is needed?

- Fundamentals theoretical background of PES (Chapter 11)
- Basics of the chosen calculation method: Hartree-Fock theory (Chapter 11).
- Fundamentals of molecular dynamics (Chapter 5)

1. Introduction

The most of methods described in the previous section are widely used in the everyday practice of the biomolecular modelling. In this chapter we give a few examples how a practical problem is solved with these methods. In the present version only two topics are discussed, but we plan to extend it with further ones from time to time.

In the first example PES-s of different protonation states of histamine is prepared. PES helps to shed light into the conformational properties of the chosen molecule and to understand the geometrical consequences of the different protonation states.

The second case study deals with the molecular dynamical refinement of an experimental polypeptide structure (Trp-cage miniprotein). The stability of its spatial structure is also characterized.

Posing the problems, the applied methods as well as the evaluation of the results are presented here without the technical details of the usage of the computer codes applied.

2. The potential energy surface of histamine

In chapter 11.4 we introduced the potential energy surface (PES) as the consequence of the decoupling between nuclei and the electrons motion or, as we can also say, the consequence of the Born-Oppenheimer approximation. The PES of a molecule is a hypersurface in a p+1 dimensional coordinate space, where p is the number of geometric parameters which characterize the geometry of the nuclei and the +1 dimension is came from the total energy of the electron system. Evidently, if we have one characteristic parameter (e.g. the angle in water molecule between the two covalent bonds), then PES would be a curve. For two parameters it is a real surface in a three dimensional coordinate system and for more parameters it is a hypersurface. Therefore, in the present case we will analyze a system with two parameters which can be represented by a surface. The subject of this case study is the histamine molecule whose importance in the human system is well known. It has many roles in many biological processes just like local immune responses or physiological function regulation. Moreover, it is also known as a neurotransmitter. From biochemical point of view, histamine is built up from two fragments: an ethanamine chain and an imidazole ring. The amino group at the end of the chain is protonated at physiological conditions while the imidazole ring can have mainly three different protonation states under the same circumstances. In *Figure 13.1*. we present these states and one can expect that such a difference would manifest itself in the potential energy surface (PES).

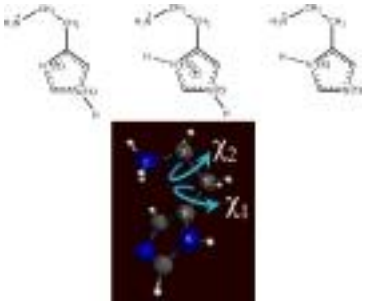
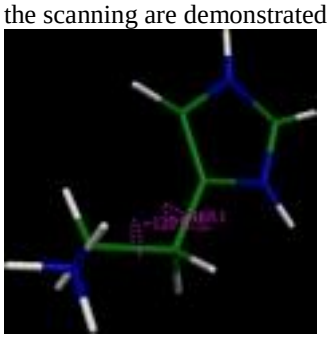


Figure 13.1. The investigated three protonation forms of histamine (τ-histamine, ττ-histamine and π-histamine, respectively) and the definition of the torsional angles with atom numbering.

To generate the PESes systematic scanning was performed with relaxation applying HF calculation method with 321 gaussian basis set. Two torsional angles characterize the PES where the rotation of the imidazole group is associated with χ1 (defined by atoms 3-5-8-9, see *Figure 13.1*.) and the rotation of the $CH_2NH_3^+$ group is described by χ2 (defined by atoms 1-3-5-8, see *Figure 13.1*.). The scanning was started from an elongated conformation of the ethanamine in each cases by setting the torsional angles to 180°. "The first few steps from the scanning are demonstrated by video 13.1."



Video 13.1. Few steps from a systematic relaxed PES scanning.

It is worth to note that this geometry is the global optimum of the ττ-histamine as one can notice it in the corresponding small figure in *Figure 13.4*. Relaxation means that partial optimization was performed at every fixed χ1 and χ2 values throughout the scannings.

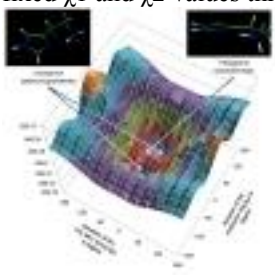


Figure 13.2. The PES of the τ-histamine with selected geometries

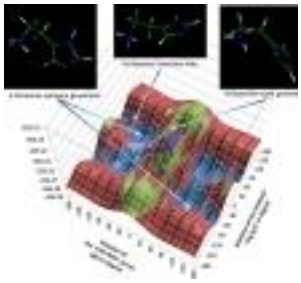


Figure 13.3. The PES of the π-histamine with selected geometries

In *Figure 13.2, 13.3.* and *13.4.* we present the results of the systematic scanning regarding the τ-, π- and ττ-histamine molecules, respectively. First, special geometries are identified on the surfaces.

The global minima of the conformers (the optimum geometries) are signed by arrows augmented with 3D picture of the minima. Because of symmetric reason there can be more than one optimum geometry but in the pictures we show only one of them. Conformers with the highest energy of the electron system ("worst geometry") are also presented in two cases together with their pictures and everybody can easily interpret which structural properties (e.g. steric hindrance, staggered and eclipsed conformations of CH_2 groups, etc.) are responsible for that disfavoured geometries.

In many cases it is also an important question how the molecule can transform from an optimum geometry to another one. One thing is sure that at least one transition state will be touched during the transformation. A transition state on the PES is a saddle point and its height with respect to the optimum geometry is related to the energy barrier between the two minima.

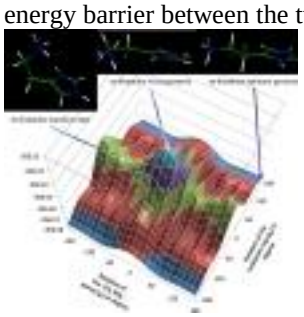


Figure 13.4. The PES of the ττ-histamine with selected geometries

In *Figure 13.2.-13.4.* we signed a few interesting transition states. It is also noteworthy that transition states are instable configurations: the molecules can be in these states only for instants but their knowledge is important for several reasons just as some of them mentioned before.

Considering the transition state geometries in pictures, it is again a routine task to find the geometric reason of their instability. We let it to the reader to find some of them.

In conclusion we can say that the differences in the protonation of the histamine are definitely expressed in the PESes, and we pointed out the importance of the selected geometries highlighted also by pictures.

Finally, we encourage everybody to go further in a more detailed investigation of the presented PES, since many more interesting information can be gained from the comparison of them (e.g. how the symmetry of the molecule is manifested on the surface; what can we learn from the comparison of the absolute total energy values, etc.).

3. Refinement and stability of protein structures: an application of MD

The optimal situation would be if a computer simulation of the folding process could be able to predict the structural properties of proteins. Unfortunately, an exclusively molecular dynamics (MD) based method works only for some small peptides having stable structures. The origins of this restricted applicability are two-folded. On the one hand the available computers and methods are unable to follow the folding process for an

appropriately long time at a proper level of accuracy. On the other hand the details of the structure formation are not completely known experimentally: *e.g.* the details of interactions with other molecules, like chaperons.

Therefore we often use initial structures obtained either from experiments (like X-ray or NMR) or from theoretical predictions (like homology modelling). However, these structures frequently need refinements, because the experimental conditions are far from the physiological ones or the quality of the homology model is not good enough for the further investigations *etc*.

In this example MD simulation of a mini protein (Trp-cage) is used for refining the structure as well as to investigate its stability. Root-mean-squared-deviation (RMSD) and Root-mean-squared-fluctuations (RMSF) of atomic coordinates are used as mathematical tools in our study.

3.1. Comparing to a reference structure

Let us suppose that we use an experimental or theoretically derived structure of a protein as a starting point in an MD simulation (with a proper solvent, temperature and appropriate co-soluted ions). During the simulation this structure goes through smaller or larger rearrangements. If the rearrangement is small, then the initial structure is close to the equilibrium structure of the protein at the simulated conditions (supposing that the force field parameterization and other parameters used provide good approximations of the physiological conditions). The large rearrangement shows the opposite. For a quantitative study we need a mathematical measure of the difference between the two structures.

3.2. RMSD,least square fitting

Let us denote the coordinates of the i-th atom of our protein $\mathbf{r}_i(t_0)$ and $\mathbf{r}_i(t)$ at the beginning (t_0) and at a time point t of the simulation. The differences of these two structures can be characterized with their root-mean-square deviations defined by the following formula

$$d_{RMSD}(t) = \left[\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{r}_i(t) - \mathbf{r}_i(t_0)\right|^2\right]^{1/2}. \qquad (13.1)$$

Here N is the number of atoms in the protein. Of course, this definition depends on the relative position and orientation of the compared geometries. However, the minimal value of d_{RMSD} is unique and can be calculated with numerical methods easily. This value is used as a quantitative measure of the "distance" of the two compared structures. The procedure used (also termed as least square fitting) provides a structural alignment where the RMSD distance of the structures from each other is minimal.

The structure of proteins is dominantly determined by the heavy (non-hydrogen atoms). The positions of H-s are less important from this point of view than other atoms in the characterization of structural similarity. Mathematically this can be formulated using the mass weighted RMSD (MW-RMSD) defined as

$$d_{MW\text{-}RMSD}(t) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i\left|\mathbf{r}_i(t) - \mathbf{r}_i(t_0)\right|^2\right]^{1/2}. \qquad (13.2)$$

where m_i is the mass of the i-th atom, $M=\Sigma_{i=1}^{N} m_i$ is the total mass of atoms in the molecule. In the protein structure characterization we not necessarily use all of the atoms, sometimes we are interested in the similarity of only a subset of atoms (*e.g.* main-chain or selected residues). In addition these subsets of atoms can be different in least square fitting and in RMSD calculation. For example if we know that the backbone remains stable during the simulation and we are interested in structural rearrangement of the side chains of certain residues than we use fitting of the backbone atoms and calculate RMSD for the selected side-chains.

If we calculate the RMSD for each residue separately we can follow the distance of a residue from its reference position along the MD trajectory and identify the time and places of the largest structural changes, too:

$$d^{R_a}_{RMSD}(t) = \left[\frac{1}{N_{R_a}}\sum_{i\in R_a}\left|\mathbf{r}_i(t) - \mathbf{r}_i(t_0)\right|^2\right]^{1/2}, \quad R_a,\ a = 1,..,N_{res} \qquad (13.3)$$

The summation is done for the N_{R_a} atoms in R_a-th residue, N_{res} is the total number of residues. The RMSD values can be calculated for a time interval of the trajectory from T_s to T for each residues (R_a, a=1,..,N_{res}) separately that gives a single number instead of a function:

$$\bar{d}^{R_a}_{RMSD} = \left[\frac{1}{T - T_s} \int_{T_s}^{T} \frac{1}{N_{R_a}} \sum_{n \in R_a} |\mathbf{r}_n(t) - \mathbf{r}_n(t_0)|^2 dt \right]^{1/2} \quad R_a, a = 1,..,N_{res} \tag{13.4}$$

In the MD calculation we have not got continuous values of coordinates. Their values are stored at K equidistant time points of the simulation. The time average for these discrete values are calculated as

$$\bar{d}^{R_a}_{RMSD} = \left[\frac{1}{K} \sum_{\tau=1}^{K} \frac{1}{N_{R_a}} \sum_{n \in R_a} |\mathbf{r}_n(t_\tau) - \mathbf{r}_n(t_0)|^2 \right]^{1/2} \quad R_a, a = 1,..,N_{res}. \tag{13.5}$$

3.3. Structural stability, RMSF

The simplest sign of the stability of a structure is that it does not change significantly during the MD simulation (equilibrium structure). But even in this case the geometry fluctuates around an average value. The extent of these fluctuations can be characterized by its average squared deviation of atomic positions from their average values, which is also called *variance* or *mean square fluctuation* (σ, MSF):

$$\sigma = \frac{1}{K} \sum_{\tau=1}^{K} \frac{1}{N} \sum_{i=1}^{N} [r_i(t_\tau) - \langle r_i \rangle]^2 \tag{13.6}$$

Here, K is the number of trajectory points and

$$\langle r_i \rangle = \frac{1}{K} \sum_{\tau=1}^{K} r_i(t_\tau) \tag{13.7}$$

is the average position of the i-th atom. The square root of σ ($\sigma^{1/2}$) is the *standard deviation* or *root-mean-squared fluctuation* (RMSF).

3.4. MD investigation of Trp-cage miniprotein

To demonstrate the ideas described above, MD simulations were performed on the Tryptophane-cage (Trp-cage) peptide (it is also called TC5B), using explicit water molecules to model the surroundings of the protein. Trp-cage is a 20-residue-long peptide with the sequence NLYIQ WLKDG GPSSG RPPPS, which is often called miniprotein, since it is one of the smallest polypeptides that possesses a well defined secondary structure [1]. The three-dimensional structural features of this peptide are well characterized by several experimental methods and molecular modelling techniques. The most important structure stabilizing factors of Trp-cage are the hydrophobic stacking of the aromatic rings of Tyr3 and Trp6 amino acids, and the salt bridge formed between the Asp9 and Arg16 residues. Its small size, temperature-sensitive structure and the simultaneous appearance of two important stabilizing interactions make this system an ideal protein model for MD studies.

The Amber _99SB-ILDN [2] force field and the TIP3P [3] water model were used in our NPT ensemble MD simulations using the GROMACS [4] molecular dynamics package.

Figure 13.5. Structure of Trp-cage (tc5b) mini-protein. The secondary structure is represented using the "new cartoon" style of VMD [5] program

Figure 13.6. presents $d_{MW\text{-}RMSD}(t)$ the values of Trp-cage miniprotein along an 50 ns long MD simulation. The backbone was fitted at each presented time point of the trajectory to those of the IL2Y structure from the Protein Data Bank (used as reference geometry). The RMSD value increases in the first 3.4 ns of the simulation and fluctuates around an average value of 1.89 Å during the remaining time.

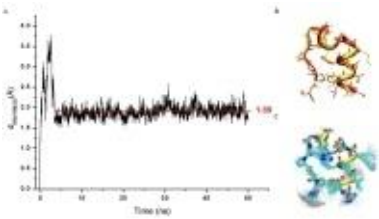


Figure 13.6. **A:** values of Trp-cage miniprotein. The average value of 1.89 Å for the last 46.6 ns is shown by the red line. **B:** The experimental (red) and average (yellow) structures of Trp-cage **C:** Structural fluctuations around the average structure: snapshots were taken from the trajectory

The average RMSD values calculated for each residue separately (black line) are shown in *Figure 13.3*. Besides the C and N terminal residues the largest deviations appears at residues Gln5, Lys8, Ser14 and Arg16. It is worth mentioning that the latest residue is part of the structure stabilizing salt bridge of Trp-cage miniprotein. The RMSD of the backbone atoms (red line) are also presented in *Figure 13.7*. These RMSD values are smaller than 1 Å for the non-terminal residues.

Figure 13.7. $d^{R}_{aMW\text{-}RMSD}$ values of Trp-cage miniprotein for the backbone atoms and for all atoms of each residues, separately. The experimental structure was used as reference.

The structural stability of the residues can be characterized by the RMSF of their atoms (*Figure 13.8*.). It is in fact calculated like the RMSD but the reference structure is the average one. The backbone fluctuations are around 0.5 Å with the exception of the terminal residues. The residues Gln5, Lys8 and Arg16 own large fluctuations together with the Pro11 and Ser12 of the central turn structure. The residue Trp6 in the centre of the hydrophobic core is stabilized by its environment as it is proved by the small RMSF values as well as the animation.

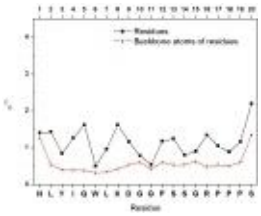


Figure 13.8. RMSF values of Trp-cage miniprotein for each residues

Figure 13.9. A short animation of the structural fluctuations around the average (yellow) structure.

4. Binding affinity estimation

The free-energy change associated with the binding of a ligand to a protein can be estimated with the Linear Interaction Energy (LIE) method [] (see Chapter 9). This method appears to be a good compromise between accuracy and computational efficiency but it does not perform equally well for all systems. The first step in its application is the determination of some system dependent parameters by using experimental binding free energy data of a set of related compounds all bound to the same protein. This fitting does not only provides us with the parameters required but is also gives us information on how well LIE works for our system.

The application presented below estimates the binding free energy between prolyl oligopeptidase (POP) and some of its ligands. It will also show that the calculations give insight into the details of the binding and the relative importance of electrostatic versus van der Waals interactions. These pieces of information may be exploited in ligand design.

POP is a serine protease that cleaves the peptide bond on the carboxy side of proline residues. A series of its inhibitors is shown in

Figure 13.10. Structure if prolyl oligopeptidase inhibitors studied. Reprinted with permission from J. Med. Chem., 51, 7514–7522, (2008). Copyright 2008 American Chemical Society.

The calculations were performed as follows. The X-ray structure of complexes **6**, **8** and **11** were used as starting points (see 13.11.). Complex structures of **5**, **7**, **9** and **10** were generated by manipulating the experimental structures (see with the Sybyl modelling suite [].

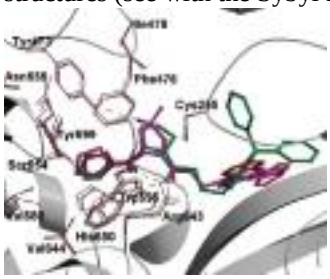


Figure 13.11. Binding of the P1-P2 moiety of the inhibitors in the crystal structures. The ribbon model of the protein is colored gray, while the inhibitor molecules and POP binding sites of the P1-P2 moieties are magenta, red and green for the POP-6, POP-8 and POP-11 complexes, respectively. Hydrogen bonds are shown as shaded lines. Reprinted with permission from J. Med. Chem., 51, 7514–7522, (2008). Copyright 2008 American Chemical Society.

Short molecular dynamics simulation were performed and the electrostatics (Eele), and van der Waals (EVDW) interaction energy components together with ligand internal energy (Econf) were extracted. (Details of the simulation protocol are described in ref .) The binding free energy was estimated from the equation

$$\Delta G=\frac{1}{2}\left(\langle E^{el} \rangle_{bound} - E^{el} \rangle_{free} \right)+\left(\langle E^{vdw} \rangle_{bound} - \langle E^{vdw} \rangle_{free} \right)+\left(\langle E^{conf} \rangle \right) \quad (13.8)$$

Note that the first two terms appear in the original LIE equation []. The first term represents the difference of the ligand-enzyme and ligand-water electrostatic interaction energies as calculated for the complex and for the solvated free ligand. The second term is similar for the van der Waals interaction energy. The last term includes the internal energies of the ligand in the bound and free states. The inclusion of this term was found to significantly improve the reproduction of experimental binding free energies. On the other hand, for the van der Waals term no multiplying factor different from 1 was found to improve the results. Experimental binding free energies were obtained from the measured ligand IC50 values (concentration causing 50% inhibition) with the formulas $\Delta G=RT\ln Ki$ and $Ki=IC50/(1+S/Km)$ [], where K_i is the inhibition constant, S is the substrate concentration and K_m is the Michaelis constant of the enzyme (see e.g. ref).

Calculated and experimental binding free energies are shown in Table 13.1 and *Figure 13.12*. below.

Table 13.1. Experimental and calculated binding free energies and their components (kcal/mol) Adapted with permission from J. Med. Chem., 51, 7514–7522, (2008). Copyright 2008 American Chemical Society.

Code	ΔE^{ele}	ΔE^{vdw}	ΔE^{conf}	ΔG^{calc}	ΔG^{exp}
5	19.4	-21.0	0.3	-10.9	-10.6

Code	ΔE^{ele}	ΔE^{vdw}	ΔE^{conf}	ΔG^{calc}	ΔG^{exp}
6	--	--	--	--	-11.8
7	18.0	-21.4	0.8	-11.6	-9.1
8	13.6	-26.5	6.1	-13.6	-11.3
9	23.0	-22.2	1.7	-9.0	-8.9
10	36.1	-26.3	6.2	-2.1	-4.0
11	20.8	-25.3	7.2	-7.7	-9.1

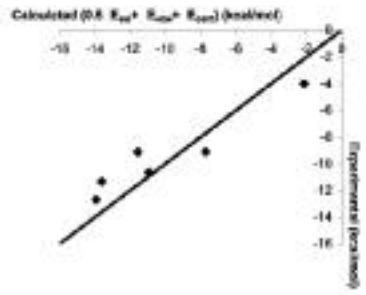


Figure 13.12. Calculated versus experimental binding free energies. The indicated line corresponds to perfect matching of the calculated and experimental values. Reprintedwith permission from J. Med. Chem., 51, 7514–7522, (2008). Copyright 2008 American Chemical Society.

The mean unsigned error is 1.37 kcal/mol and this is remarkably good taking into account the simplicity of the model applied. A favorable van der Waals and an unfavorable electrostatic energy change accompanied binding for all molecules studied. The pro-pyrrolydine moiety adopts similar conformation in all complexes and this conformation is close to that observed in water. The conformational strain found to be significant for compounds with large N-terminal groups, namely for compounds **8**, **10** and **11**. Interestingly compound **9** has a lower conformational strain than does 8, although they have the same bulky N-terminal group. This can be rationalized by the longer alkyl chain of the former that gives more flexibility to the molecule.

Further analysis of the binding modes and their energetic consequences can be found in ref. .

5. Summary

In this chapter we gave examples of the application of some methods of biomolecular modelling on practical problems. Namely, molecular dynamics and a quantum chemical method were used for obtaining structural information for molecules with biological importance.

6. References

1. J.W. Neidigh, R.M. Fesinmeyer, and N.H. AnderNat. Struct. Biol. 9, 425 (2002).

2. K. Lindor-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, and D.E. Shaw. Proteins 78, 1950 (2010).

3. W.L. Jorgensen, J. Chandrasekhar, J. Madura and M.L.Klein, J. Chem. Phys. 79, 926 (1983).

4. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. 4, 435 (2008); www.gromacs.org

5. http://www.ks.uiuc.edu/Research/vmd/

6. J. Åqvist, C. Medina, J.–E. Samuelsson, “A new method for predicting binding affinity in computer-aided drug design”, Protein Eng., 7, 385-391, (1994).

7. K. Kánai, P. Arányi, Z Böcskei, G. Ferenczy, V. Harmat, K. Simon, S. Bátori, G. Náray-Szabó, I. Hermecz “Prolyl Oligopeptidase Inhibition by N-Acyl-pro-pyrrolidine-type Molecules” J. Med. Chem., 51, 7514–7522, (2008).

8. SYBYL Version 6.7, Tripos Inc., St. Louis, MO.

9. Cheng, Y.-C.; Prusoff, W.H. “Relationship between inhibition constant (Ki) and the concentration of inhibition which causes 50% inhibition (IC50) of and enzyme reaction.” Biochem. Pharmacol. 22, 3099-3108, (1973).

10. J.M. Berg, J.L. Tymoczko, L. Stryer “Biochemistry”, Freeman, New York, 2002.

7. Questions

1. What is the meaning and mathematical definition of RMSD and RMSF?
2. How can the more and less stable residues be identified using the RMSF values obtained from an MD trajectory?
3. How is the symmetry of a molecule manifested on the different PES-s?
4. What can we learn from the comparison of the absolute values of total energy at the different PES-s?
5. How can we identify mathematically a minimum or saddle point on a PES?
6. Please choose a stationary point from the presented PES-s and discuss the distinct structural properties of the corresponding molecular conformation (e.g. steric hindrance, staggered and eclipsed conformations of CH2 groups, etc.).

8. Glossary

- *RMSD*: Root-mean-squared-deviation
- *RMSF*: Root-mean-squared-fluctuation