Lecture 8: Protein Modeling

Junmei Wang

Department of Pharmacology, University of Texas Southwestern Medical Center at Dallas

Junmei.wang@utsouthwestern.edu

Project 2: MD Simulations

1. Select a protein system

Protein Class	PDB Code	-logk_d	Resolution
Neuraminidase	2QWG	8.4	1.8
DHFR	1DHF	7.4	2.3
L-arabinose	1ABE	6.52	1.7
Thrombin	1A5G	10.15	2.06
Human oxresin receptor 1	4ZJ8	~10	2.75

Project 2 MD Simulations -Continued

2. Select top hits from autodock-vina screening
   - Top 2 and bottom 1

3. Prepare ligand structures and residue topologies
   - Add hydrogen with adt
   - Generate Gaussian gcrt file with antechamber
   - Run G09 to calculate electrostatic potentials (ESP)
   - Run Antechamber to assign RESP charges
   - An alternative is run antechamber to assign am1-bcc charge

Project 2 MD Simulations -Continued

4. Prepare topology files for minimization and MD simulations
   - xleap
   - tleap

5. Run Minimization and MD simulations using a delicate scheme
   - Minimization with main chain restrained using a set of gradually reduced restraint force constants
   - MD simulation with main chain restrained using a set of gradually reduced restraint force constants
   - Heat systems up using a set of temperatures
   - Equilibrium phase
   - Sampling phase

Project 2 MD Simulations -Continued

6. Analyze MD snapshots
    - Average structure
    - RMSD ~ simulation time plots
    - Quasi-harmonic analysis
    - MD movie

Project 3: Binding Free Energy Calculations With MM-PB/GBSA

1. Select a protein system

Protein Class	PDB Code	-logk$_d$	Resolution
Neuraminidase	2QWG	8.4	1.8
DHFR	1DHF	7.4	2.3
L-arabinose	1ABE	6.52	1.7
Thrombin	1A5G	10.15	2.06
Human oxresin receptor 1	4ZJ8	~10	2.75

Project 3: Binding Free Energy Calculations With MM-PB/GBSA

2. Prepare topologies for energy calculations with implicit solvent
   - Xleap
   - Tleap

3. Run mmpbsa.py to do the calculation
   - Input file
   - Output files

4. Analyze the MM-PB/GBSA results

Protein Modeling

Contents

- Introduce the process of homology modelling.
- Summarise the methods for predicting the structure from sequence.
- Describe the individual steps involved in creating and optimising a protein homology model.
- Outline the methods available to evaluate the quality of homology models.
- Case Study – Modelling the Drug binding site of hERG.

Why Homology Model?

- Solving protein structures is not trivial.
- There are currently ~**1.8 million** known protein coding sequences.
- But only ~**120,000** protein structures in the PDB.
- Even so, many of these structures are duplicates.
- For Membrane Proteins structural data is even more sparse:
- There are currently **2829** membrane protein structures

RSCB Protein Data Bank (PDB)
Statistics (22/07/16)

Method	**Totals**
X-ray	117790
NMR	11469
EM	1089
Other	198
Total	**120642**

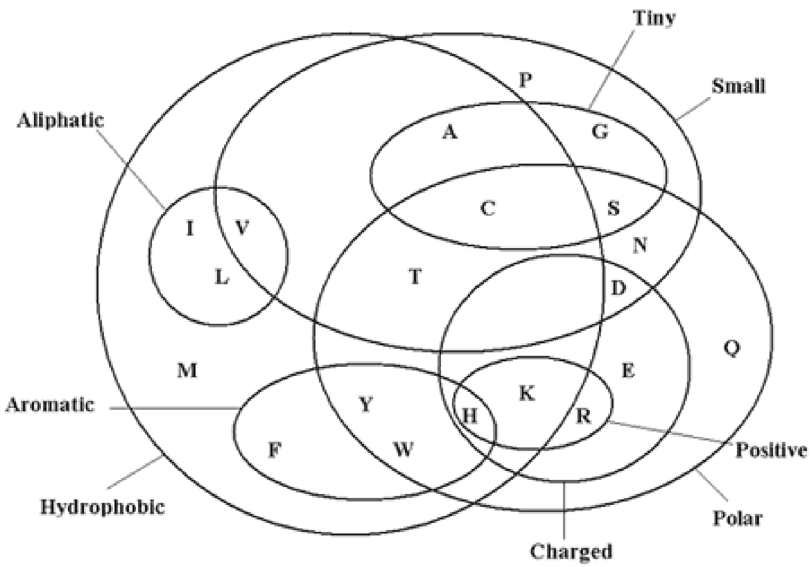
www.rscb.org

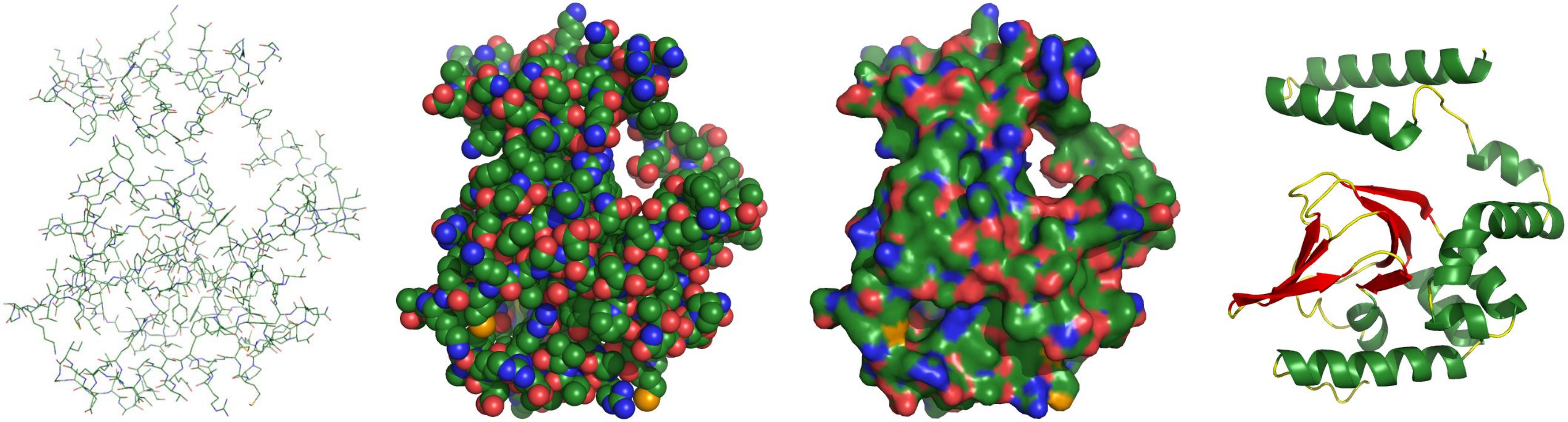
Amino Acid Residues

- Proteins are made up of amino acids, which are interconnected by peptide bonds.

- There are 20 naturally occurring amino acids.

- Amino acids may be subdivided by their individual properties.

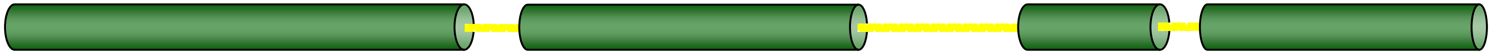




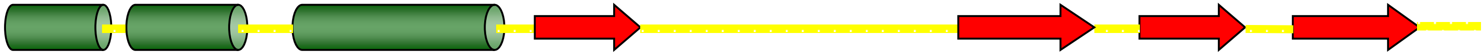
From Sequence to Structure

Primary Structure - Amino Acid Sequence

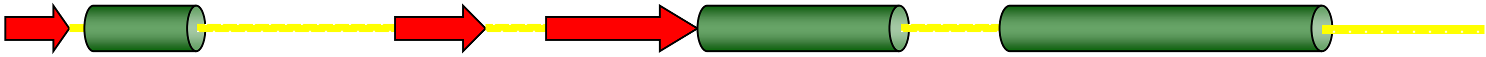
DSSRRQYQEKYKQVEQYMSFHKLPADFRQKIHDYYEHRYQGKMFDEDSILGELNGPLREEIVNFNCR



KLVASMPLFANADPNFVTAMLTKLKFEVFQPGDYIIREGTIGKKMYFIQHGVVSVLTGNKEMKLSDG



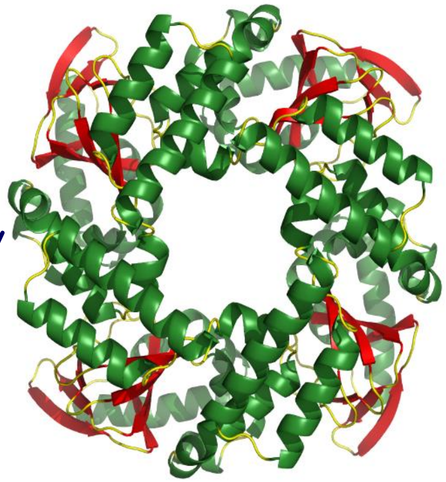
SYFGEICLLTRGRRTASVRADTYCRLYSLSVDNFNEVLEEYPMMRRAFETYVAIDRLDRIGKKNSIL



Secondary Structure

Tertiary Structure

Quaternary Structure



What information can we get from a Sequence of amino acids?

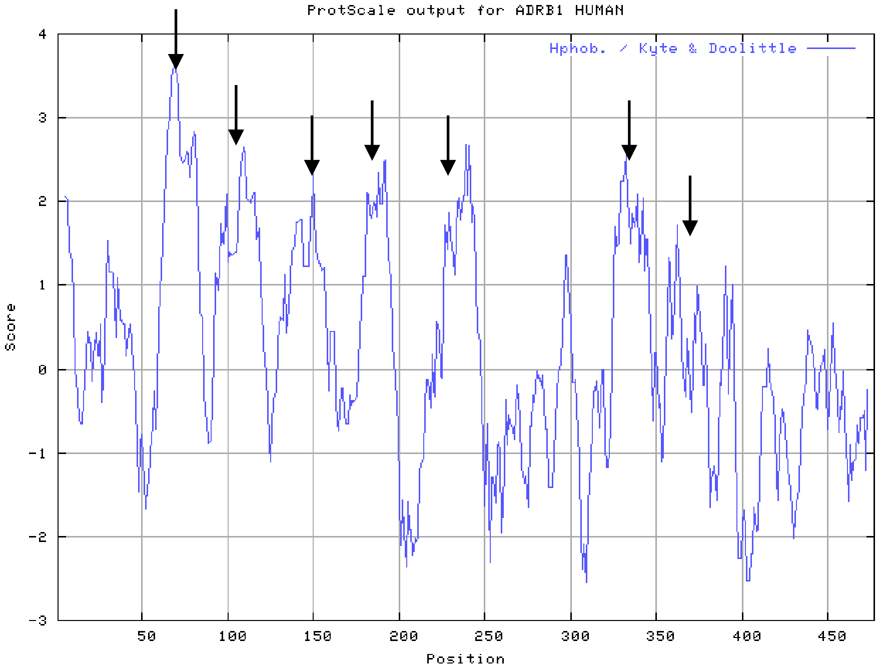
Secondary Structure Prediction

- The **S**econdary **S**tructure of **P**roteins is **D**efined by the **DSSP** algorithm.
- Amino acids classified as either α-helix (H), β-strand (S) or loop (C).
- It is possible to extract structural information from amino acid sequence.
- These prediction methods were initially proposed by Chou & Fasman in 1978.
- They used a statistical method based on 15 known crystal structures.
- Recent developments and an increase in structural information has improved these methods and they are currently ~80% accurate.

PSI-Pred:	http://bioinf.cs.ucl.ac.uk/psipred/
JPred:	http://www.compbio.dundee.ac.uk/~www-jpred/

Transmembrane Helix Prediction

- The amino acids at the centre of transmembrane helices are generally hydrophobic in nature.
- Analysis of Hydropathicity can be used to predict the number of membrane spanning helices.
- The analysis for the G-protein coupled receptor to the right suggests it has 7 TM helices.
- The example used the Kyte & Doolittle scale.



Hydropathy Plot

http://expasy.org/tools/protscale.html

BLAST

- How to find an appropriate template Structure for homology modelling…

- **B**asic **L**ocal **A**lignment **S**earch **T**ool

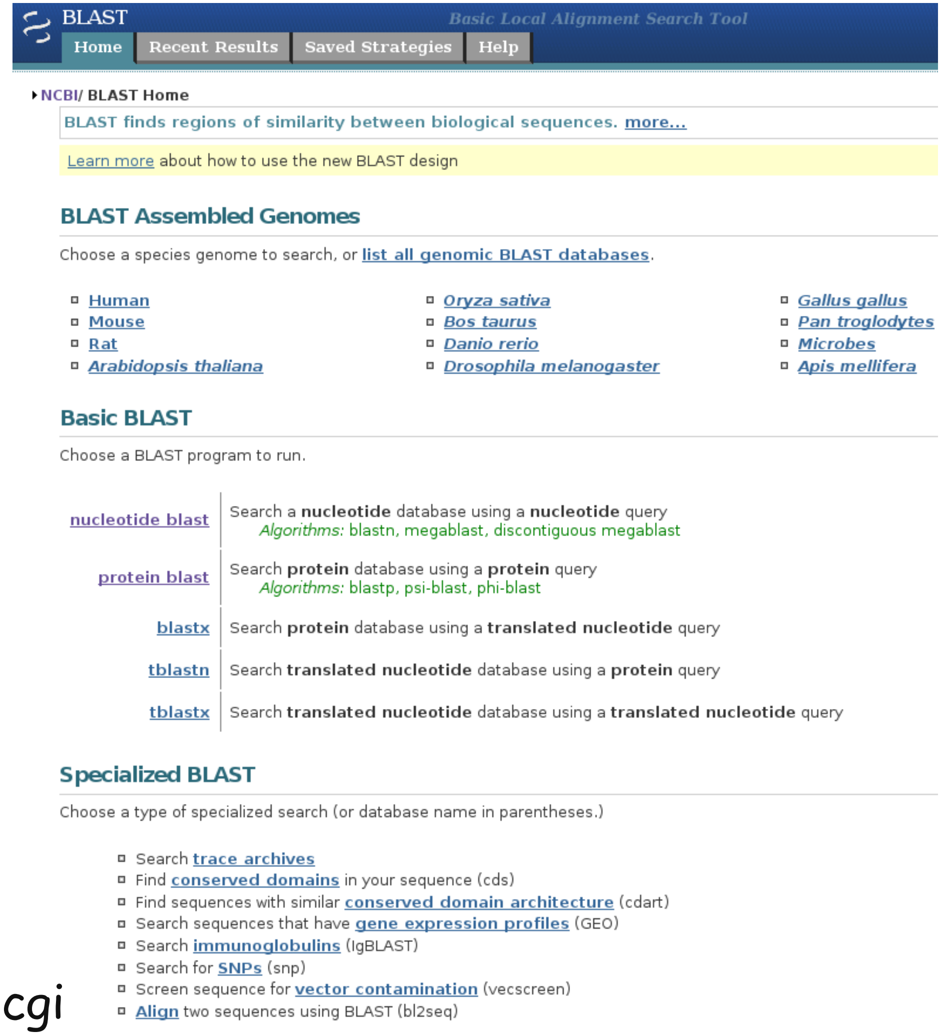
- Used to search protein databases:

- e.g. Non-redundant (nr) & SwissProt to find similar **sequences**.

- Protein Data Bank (PDB) to find **structures** with similar sequences.

- PSI- & PHI-blast are more advanced Blast methods.

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

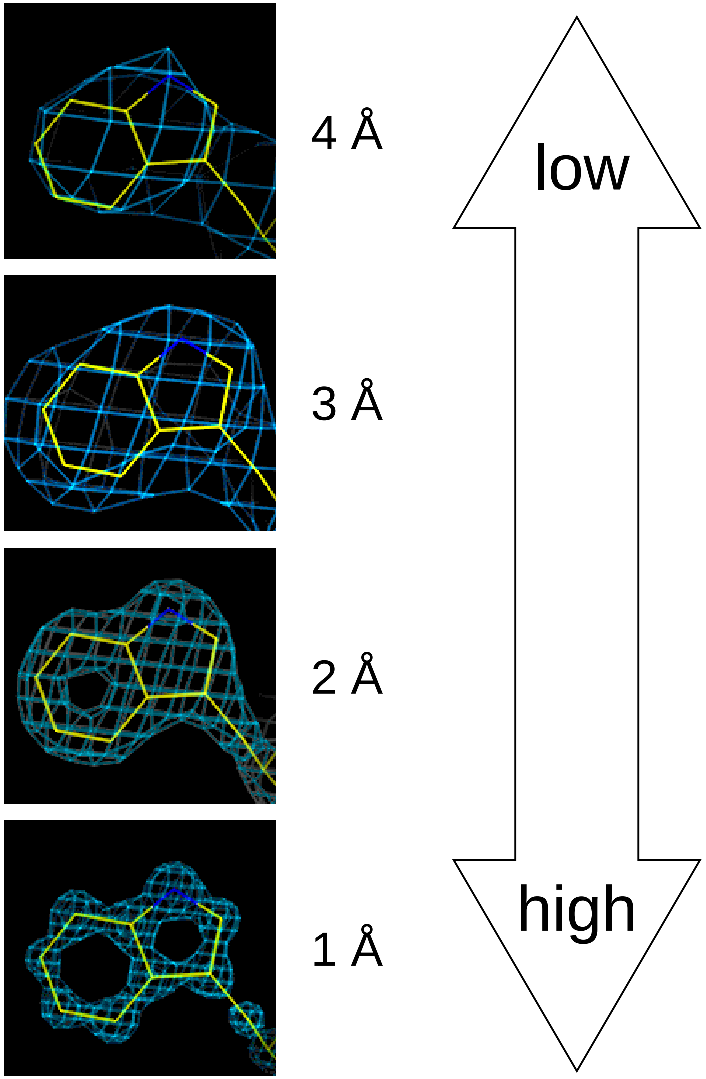
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two sequences using BLAST (bl2seq)

The Importance of Resolution



- In X-ray crystallography it is not always possible to flawlessly resolve the crystal density of the protein of interest.

- This results in a lower resolution structure.

- The lower the resolution the more likely the structure is wrong.

- The resolution of the template structure also reflects in the quality of the homology model.

Sequence Alignment

Aligns the sequence(s) of interest to that of the template structure(s):

- **Emboss** may be used for two sequence, to generate a pairwise alignment & a percentage identity – ideally an identity of >50%:

  http://www.ebi.ac.uk/emboss/align/

- **T-Coffee**, **Clustal** & **MUSCLE** are popular methods for multiple sequence alignment. All may be found at :

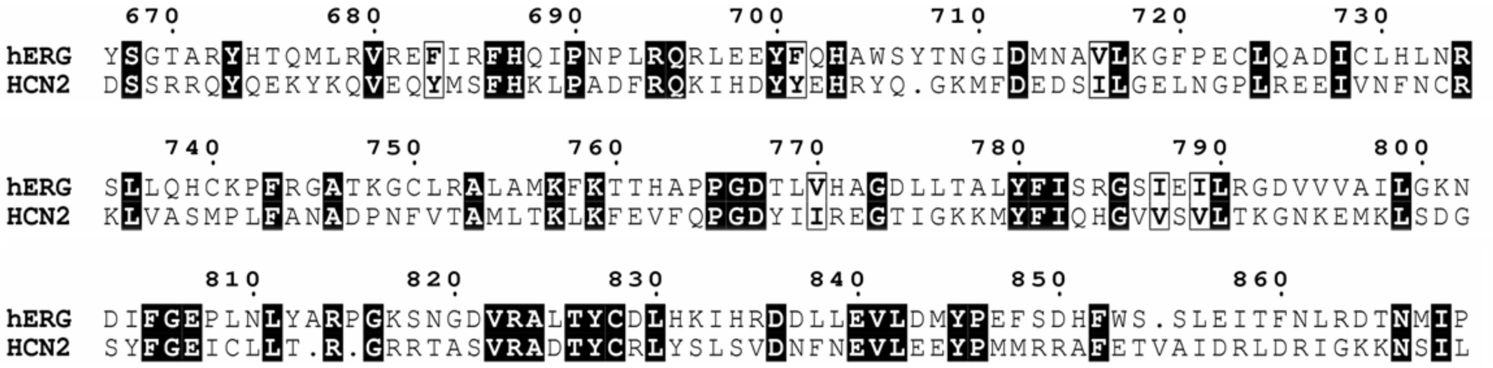
  http://www.ebi.ac.uk/

- **ESPRIPT** is useful for formatting to creating black & white figures:

  http://espript.ibcp.fr/

- **Promals3d**: Nick Grishin at UTSW

  http://prodata.swmed.edu/promals3d/promals3d.php



```
           670       680       690       700       710       720       730
hERG    YSGTARYHTQMLRVREFIRFHQIPNPLRQRLEEYFQHAWSYTNGIDMNAVLKGFPECLQADICLHLNR
HCN2    DSSRRQYQEKYKQVEQYMSFHKLPADFRQKIHDYYEHRYQ.GKMFDEDSILGELNGPLREEIVNFNCR

             740       750       760       770       780       790       800
hERG    SLLQHCKPFRGATKGCLRALAMKFKTTHAPPGDTLVHAGDLLTALYFISRGSIEILRGDVVVAILGKN
HCN2    KLVASMPLFANADPNFVTAMLTKLKFEVFQPGDYIIREGTIGKKMYFIQHGVVSVLTKGNKEMKLSDG

               810       820       830       840       850          860
hERG    DIFGEPLNLYARPGKSNGDVRALTYCDLHKIHRDDLLEVLDMYPEFSDHFWS.SLEITFNLRDTNMIP
HCN2    SYFGEICLLT.R.GRRTASVRADTYCRLYSLSVDNFNEVLEEYPMMRRAFETVAIDRLDRIGKKNSIL
```

Automated Homology Modelling

If you are *lazy* there are servers that do the modelling for you!

- Swiss Model : http://swissmodel.expasy.org//SWISS-MODEL.html
- Robetta : http://robetta.bakerlab.org/
- 3D Jigsaw : http://www.bmm.icnet.uk/servers/3djigsaw/
- Phyre : http://www.sbg.bio.ic.ac.uk/phyre/
- EsyPred3D : http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
- CPHmodels : http://www.cbs.dtu.dk/services/CPHmodels/

Eva-CM performs continuous and automated analysis of comparative protein structure modeling servers
http://pdg.cnb.uam.es/eva/doc/intro_cm.html

Modeller

- Well regarded program for Homology/Comparative Modelling.
- Current Version 9.17 https://salilab.org/modeller/
- Requires an Input file, Sequence alignment & Template structure.

```
from modeller import *
from modeller.automodel import *
log.verbose()
env = environ()
env.io.atom_files_directory = './'

a = automodel(
  env,
  alnfile = 'herg.ali',
  knowns = '1q5o',
  sequence = 'herg'
)

a.starting_model= 1
a.ending_model  = 1
a.make()
```

Input File (*.py)

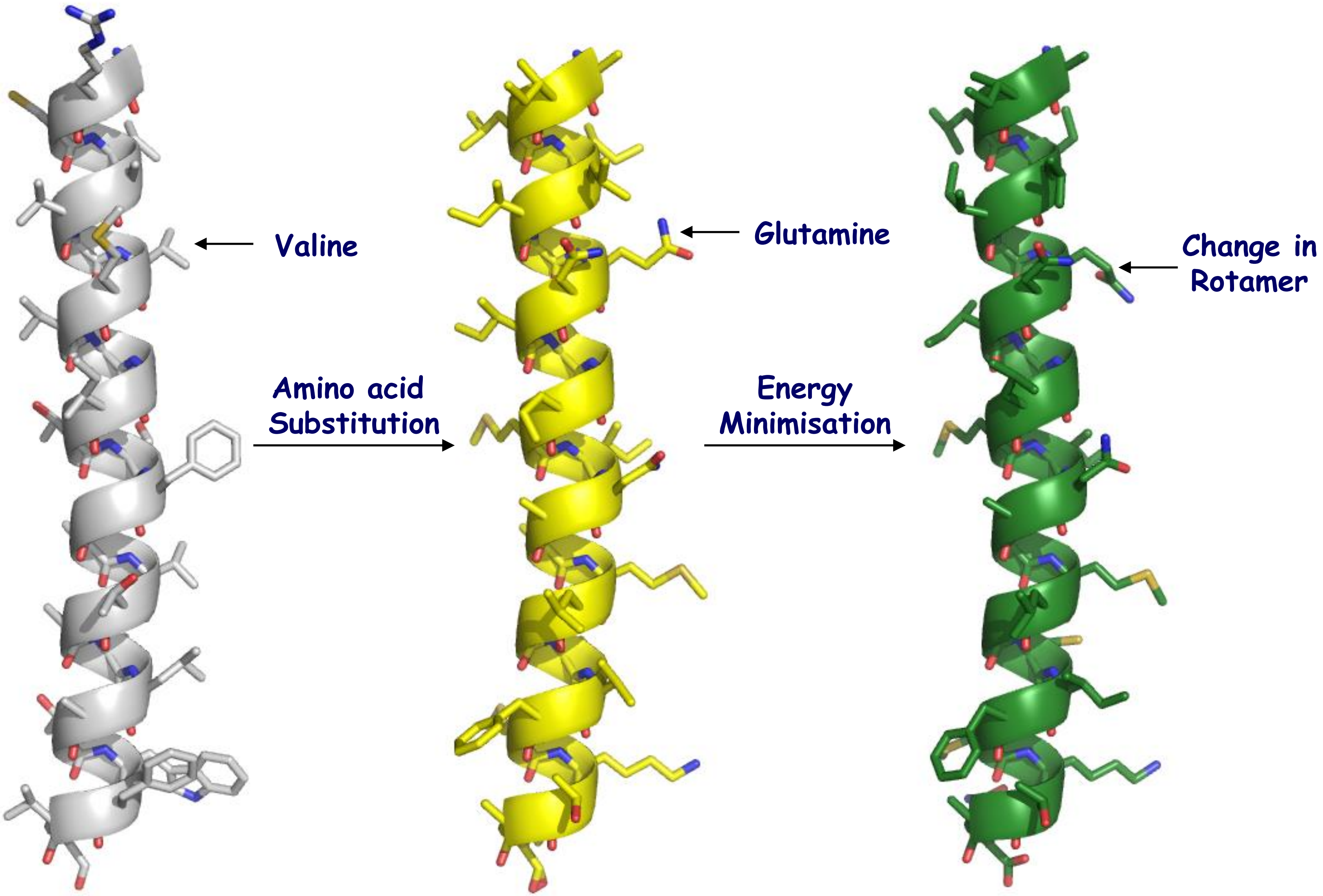
```
>P1;1q5o
structureX: 1q5o : 443 : A : 644 : A ::::
DSSRRQYQEKYKQVEQYMSFHKLPADFRQKIHDYYEHRYQ-GKMFDEDSILGELNGPLRE
EIVNFNCRKLVASMPLFANADPNFVTAMLTKLKFEVFQPGDYIIREGTIGKKMYFIQHGV
VSVLTKGNKEMKLSDGSYFGEICLL--TRGRRTASVRADTYCRLYSLSVDNFNEVLEEYP
MMRRAFETVAIDRLDRIGKKNSIL.*

>P1;herg
sequence: herg : 1 :::::::
YSGTARYHTQMLRVREFIRFHQIPNPLRQRLEEYFQHAWSYTNGIDMNAVLKGFPECLQA
DICLHLNRSLLQHCKPFRGATKGCLRALAMKFKTTHAPPGDTLVHAGDLLTALYFISRGS
IEILRGDVVVAILGKNDIFGEPLNLYARPGKSNGDVRALTYCDLHKIHRDDLLEVLDMYP
EFSDHFWSSLEITFNLRDTN-MIP.*
```

Sequence Alignment (*.ali)

```
ATOM   1   N    ASP A 443   -15.943  41.425  44.702  1.00 44.68
ATOM   2   CA   ASP A 443   -15.424  42.618  45.447  1.00 43.15
ATOM   3   C    ASP A 443   -14.310  43.306  44.686  1.00 41.81
ATOM   4   O    ASP A 443   -14.298  44.528  44.539  1.00 42.61
                                  etc...
```

Template Structure (*.pdb)

How Does it Work?



Valine

Amino acid Substitution

Glutamine

Energy Minimisation

Change in Rotamer

Template Structure

Initial Model (*.ini)

Output Model(s) (*.B999*)

Modeller : Output

- **.log** : log output from the run.
- **.B*** : model generated in the PDB format.
- **.D*** : progress of optimisation.
- **.V*** : violation profile.
- **.ini** : initial model that is generated.
- **.rsr** : restraints in user format.
- **.sch** : schedule file for the optimisation process.

Modeller Features & Restraints

- **Secondary Structure.**
  Regions of the protein may be forced to be α-helical or β-strand.

- **Distance restraints.**
  The distance between atoms may be restrained.

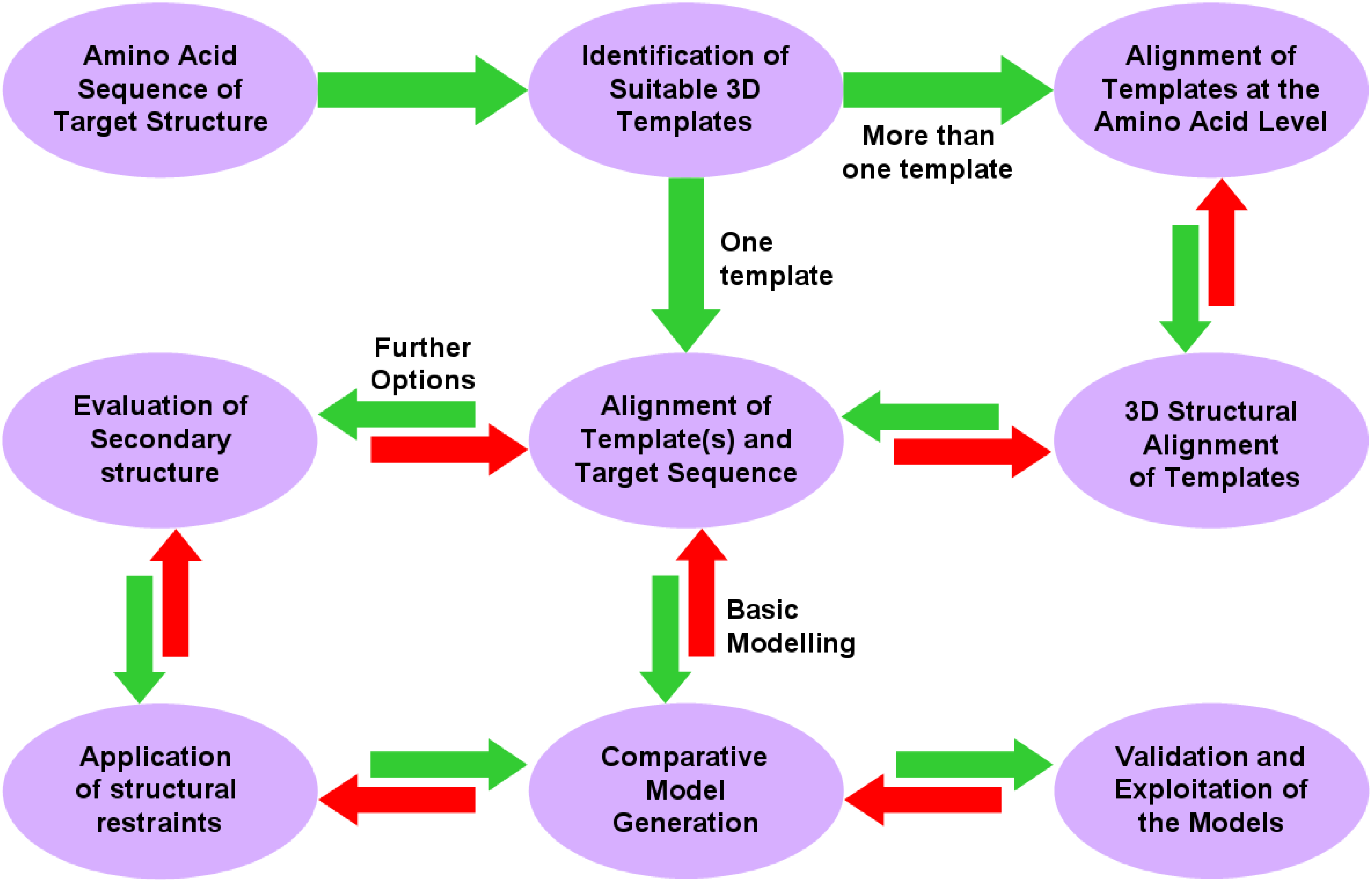
- **Symmetry.**
  Protein multimers can be restrained so that all monomers are identical.

- **Disulphide Bridges.**
  Two cysteine residues in the model can be forced to make a cystine bond.

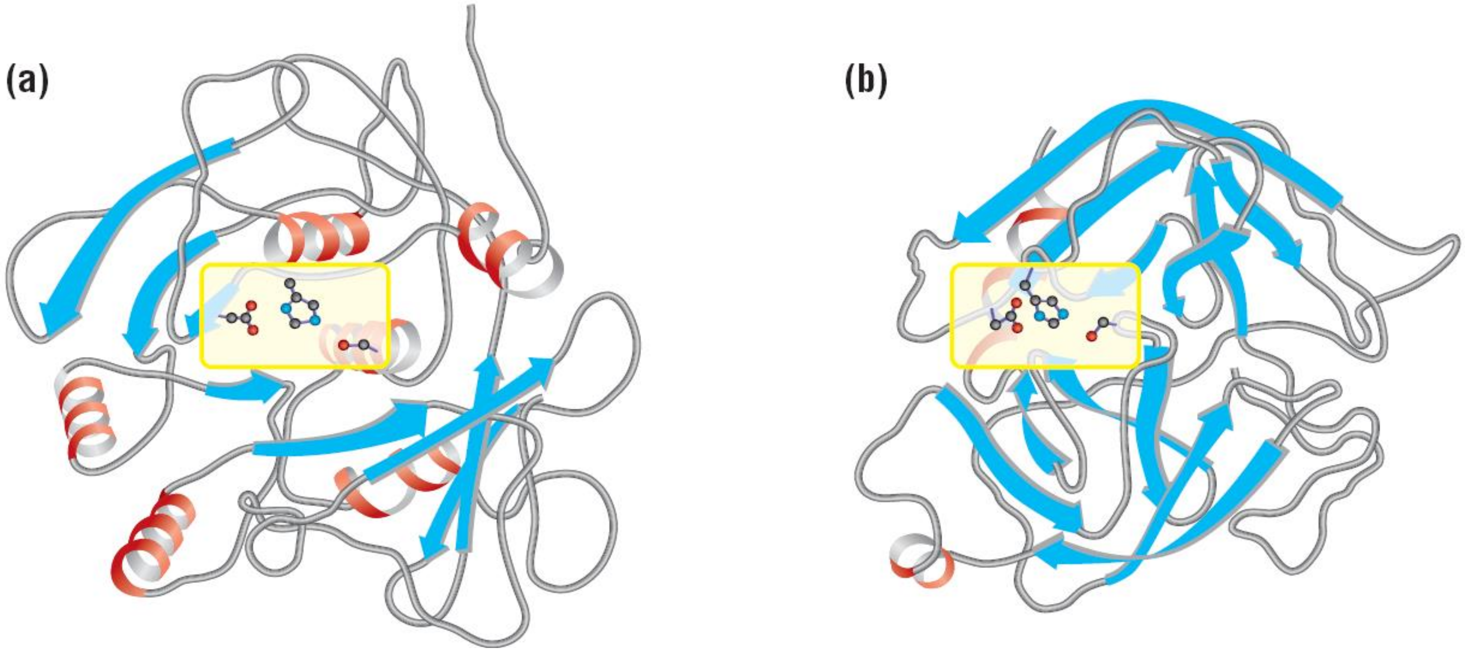
- **Ligands.**
  Ions, waters and small molecules may be included from the template.

- **Loop Refinement.**
  Regions without secondary structure often require further refinement.

An Iterative Process


Amino Acid Sequence of Target Structure
Identification of Suitable 3D Templates
Alignment of Templates at the Amino Acid Level
More than one template
One template
Further Options
Evaluation of Secondary structure
Alignment of Template(s) and Target Sequence
3D Structural Alignment of Templates
Basic Modelling
Application of structural restraints
Comparative Model Generation
Validation and Exploitation of the Models

Structural Convergence



- The catalytic triad of Serine, Aspartate and Histidine is found in certain protease enzymes. **(a)** Subtilisin **(b)** Chymotrypsin.

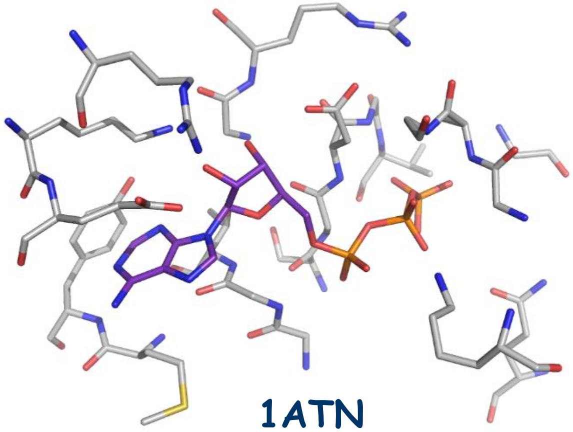
- However, the overall structure of the enzyme is often different.

- This is also important when considering ligand binding sites.

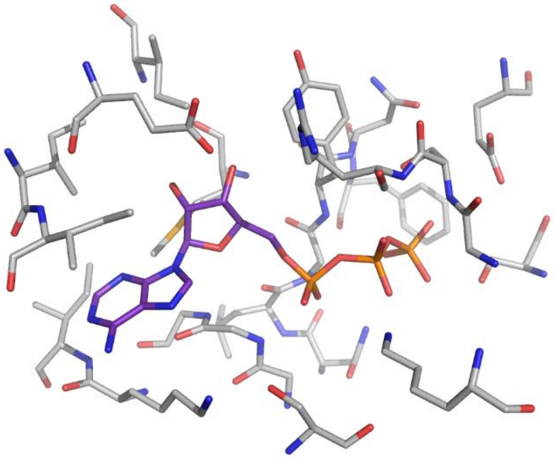
Modelling Ligand Interactions

- Small molecules, waters and ions can be retained from the template structure.
- It is possible to search for homologues based on the ligands they bind.
- Experimental data, especially mutagenesis is very useful when modelling ligand binding sites.
- Although the key residues may often remain, the overall structure of the protein may vary radically.
- The presence of the ligand is also likely to alter the conformation of the protein.

<u>ATP Binding Site</u>



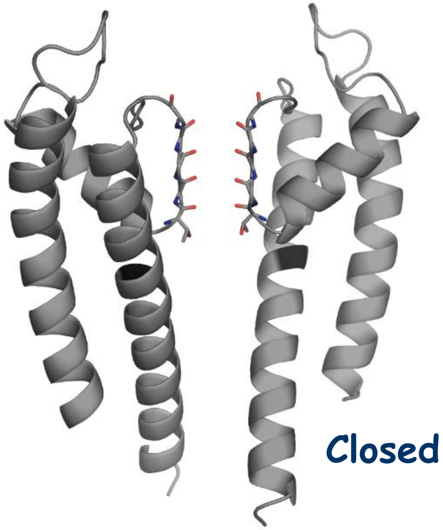
1ATN

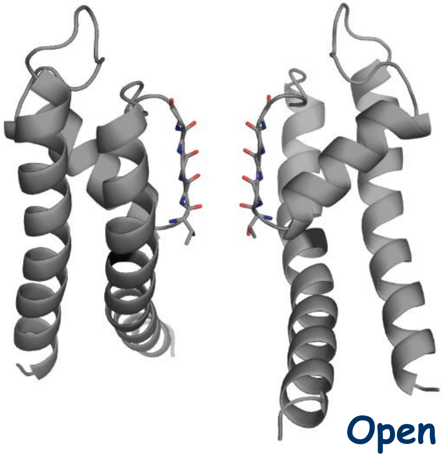


1E4G

Conformational States

- The backbone structure of the model will be almost identical to that of the template.
- Therefore the conformational state of the template will be retained in the resultant homology model.
- This is important when considering the open or closed conformation of a channel…
- … or the Apo versus bound state of a ligand binding site.





Loop Modeling

Loop Modelling

Issues with Loop Modelling

- As loops are less restrained by hydrogen bonding networks they often have increased flexibility and therefore are less well defined.
- In addition the increased mobility make looped regions more difficult to structurally resolve.
- Proteins are often poorly conserved in loop regions.
- There are usually residue insertions or deletions within loops.
- Proline and Glycine resides are often found in loops – we'll come back to this when discussing Model evaluation protocols.

Loop Modelling

- There are two main methods for modelling loops:

  1. **Knowledge based**: A PDB search for fragments that match the sequence to be modelled (Levitt, Holm, Baker etc.).

  2. **Ab initio**: A first principles approach to predict the fold of the loop, followed by minimisation steps.

- Many of the newer loop prediction methods use a combination of the two methods.

- These approaches are being developed into methods for computationally predicting the tertiary structure of proteins. eg Rosetta.

- But this is computationally expensive.

- Modeller creates an energy function to evaluate the loop's quality.

- The function is then minimised by Monte Carlo (sampling), Conjugate Gradients (CG) or molecular dynamics (MD) techniques.

Loops – the Rosetta Method

- Find fragments (10 per amino acid) with the same sequence and secondary structure profile as the query sequence.

- Combine them using a Monte Carlo scheme to build the loop.

David Baker *et al.*

Predicting Sidechain Conformations

1. Networks of side chain contacts are important for retaining protein structure.

2. Sidechains may adopt a variety of different conformations, but this is dependent on the residue type.

   For example a threonine generally adopts 3 conformations, whilst a lysine may adopt up to 81.

3. This is dependent backbone conformation of the residue.

4. The different residue conformations are known as rotamers.

5. Where a residue is conserved it is best to keep the side chain rotamer from the template than predict a new one.

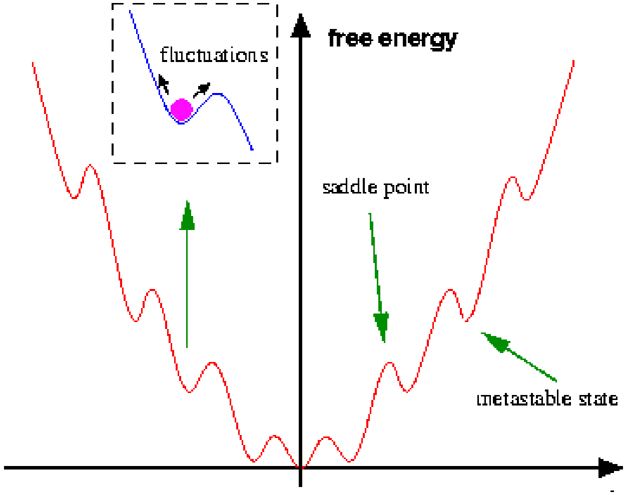
6. Rotamer prediction accuracy is high for buried residues, but much lower for surface residues:
   - Side chains at the surface are more flexible.
   - Hydrophobic packing in the core is easier to handle than the electrostatic interactions with water molecules. (cytoplasmic proteins)

7. Most successful method is SCWRL by Dunbrack *et al.*: http://dunbrack.fccc.edu/SCWRL3.php

Model Refinement

Refinement

- Energy minimization
- Molecular dynamics
    - *Big errors like atom clashes can be removed, but force fields are not perfect and small errors will also be introduced – keep minimization to a minimum or matters will only get worse.*



Error Recovery

- If errors are introduced in the model, they normally can NOT be recovered at a later step
  - The alignment can not make up for a bad choice of template.
  - Loop modeling can not make up for a poor alignment.
- If errors are discovered, the step where they were introduced should be redone.

Model Validation

Validation

1. Stereochemical checks on bond lengths, angles and atomic contacts
   Most programs will get the bond lengths and angles right

2. Ensures that the backbone conformation of the model is normal.

3. Evaluate the Ramachandran Plot
   The Ramachandran plot of the model usually looks pretty much like the Ramachandran plot of the template

4. Check the inside/outside distributions of polar and apolar residues

5. Check if validate the known biological/biochemical data
   - Active site residues
   - Modification sites
   - Interaction sites

Model Evaluation With Modeller

1. For every model, Modeller creates an objective function energy term, which is reported in the second line of the model PDB file (*.B**).

   This is not an absolute measure but can be used to rank models calculated from the same alignment. The lower the value the better.

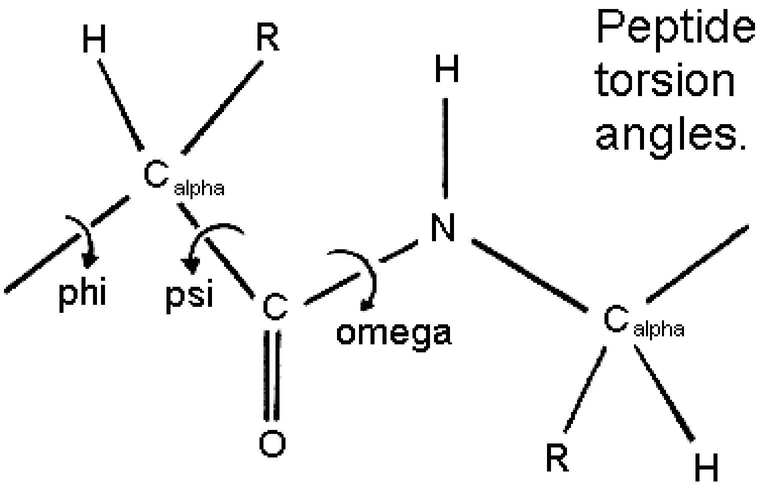
2. DOPE scoring

3. A Ca-RMSD (Root Mean Standard Deviation) between the template structure and models can also be used to compare the final model to its template.

   A good Ca-RMSD will be less than 2Å.

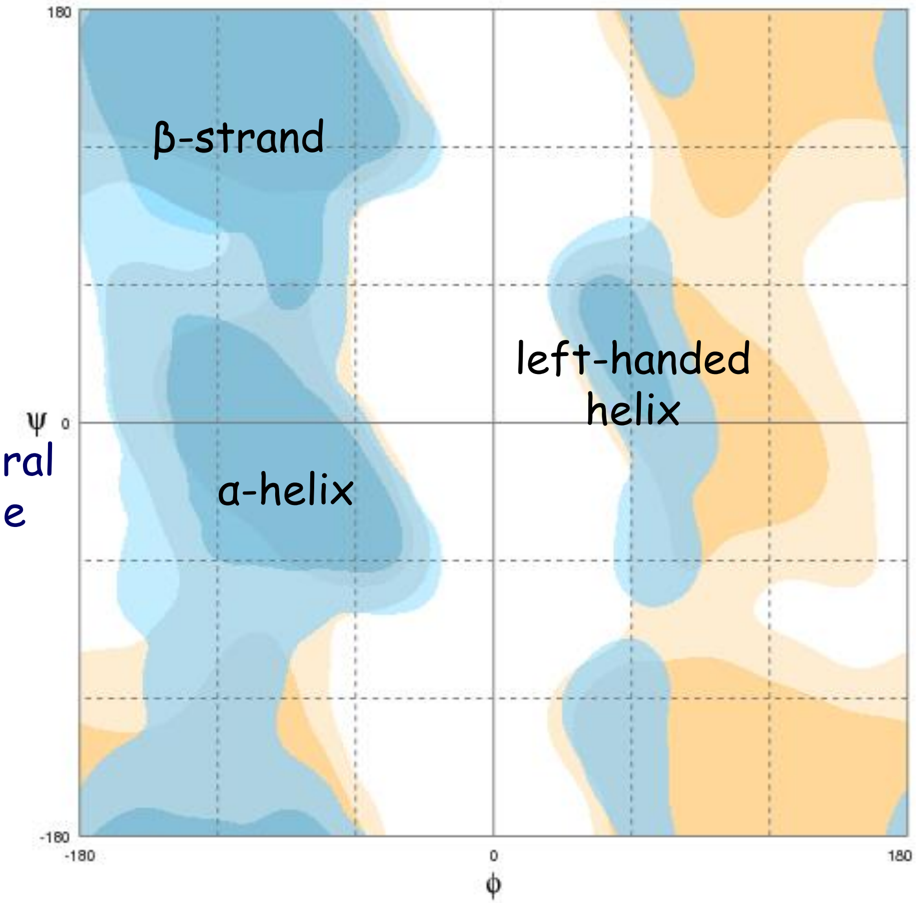
4. Modeller is good on the whole, but sometimes struggles with residues found in loops.

Ramachandran Plot



Peptide dihedral angles

Psi Dihedral Angle

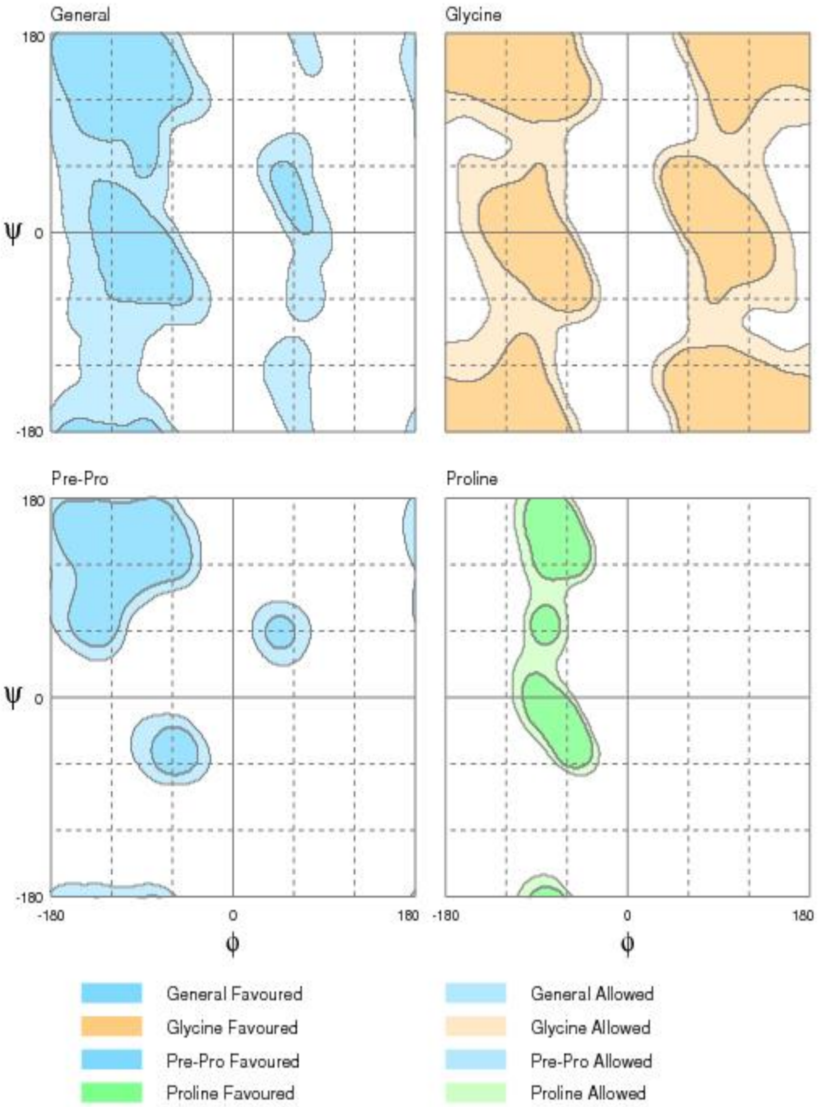


Phi Dihedral Angle

Ramachandran Plot

- The results of the ramachandran plot will be very similar to that of the template.
- A Good template is therefore key!
- Most residues are mainly found on the left-hand side of the plot.
- Glycine is found more randomly within plot (orange), due to its small sidechain (H) preventing clashes with its backbone.
- Proline can only adopt a Phi angle of ~-60° (green) due to its sidechain.

•N This also restricts the conformational space of the pre-proline residue.



Structure Validation

- ProCheck:
  http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
  http://services.mbi.ucla.edu/PROCHECK/
- WhatIf server :
  http://swift.cmbi.kun.nl/WIWWWI/
- ProQ :
  http://www.sbc.su.se/~bjorn/ProQ
- Biotech Validation Suite:
  http://biotech.embl-ebi.ac.uk:8400/
- RAMPAGE:
  http://mordred.bioc.cam.ac.uk/~rapper/rampage.php

PROCHECK

```
 +----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
 |                                                                            |
 | mgirk .pdb   2.5                                              104 residues |
 |                                                                            |
*| Ramachandran plot:   91.7% core    7.6% allow    0.3% gener    0.4% disall |
 |                                                                            |
*| All Ramachandrans:   15 labelled residues  Backbone                        |
*| Chi1-chi2 plots:      6 labelled residues  Sidechain                       |
 | Main-chain params:     6 better     0 inside      0 worse                   |
 | Side-chain params:     5 better     0 inside      0 worse                   |
 |                                                                            |
*| Residue properties: Max.deviation:    16.1              Bad contacts:   10 |
*|                     Bond len/angle:    8.0    Morris et al class:  1  1  3 |
 |                                                                            |
 | G-factors           Dihedrals:   0.10  Covalent:   0.29    Overall:   0.16 |
 |                                                                            |
 | M/c bond lengths: 99.1% within limits   0.9% highlighted                   |
*| M/c bond angles:  98.1% within limits   1.9% highlighted                   |
 | Planar groups:   100.0% within limits   0.0% highlighted                   |
 |                                                                            |
 +----------------------------------------------------------------------------+
   + May be worth investigating further.   * Worth investigating further.
```

Validation – ProQ Server

- ProQ is a neural network based predictor that based on a number of structural features predicts the quality of a protein model.

- ProQ is optimized to find correct models in contrast to other methods which are optimized to find native structures.

Arne Elofssons group: http://www.sbc.su.se/~bjorn/ProQ/

CASP

- Critical Assessment of Structure Prediction.
- A Biennial *competition* that has run since 1994.
- The next competition will be in 2008 (CASP8)
- http://predictioncenter.org/

- Its goal is to advance the methods for predicting protein structure from sequence.
- Protein structures yet to be published are used as blind targets for the prediction methods, with only sequence information released.
- Competitors may use Homology Modelling, Fold recognition or Ab Initio structural prediction methods to propose the structure of the protein.

Summary

- Homology Modelling is a valuable tool for structural biologists. It is important to take time when constructing a model.

- There are five main stages:
  1. Identify an appropriate template structure(s).
  2. Create a Sequence alignment.
  3. Perform the homology modelling.
  4. Analyse and Evaluate the quality of the model.
  5. Refinement.

- Successful homology modelling depends on the following:
  - Template quality
  - Alignment (add biological information)
  - Modelling program/procedure (use more than one)

- Always validate your final model!

Modeller: G5G8

Protein Modeling With Modeller: An Example

- **Alignment**
  1. Promals3d Alignment
  2. Generate alignment file for Modeller

- **Python script for running Modeller**
  1. Model_generation.py

- **Model selection**
  1. DOPE score
  2. GA341 score

Rosetta: PDZ3

Rosetta Protein Design (Rosetta 3.1 and up)

- https://www.rosettacommons.org/

- **Basic procedure**
  1. Generate profile for the protein to be designed. (make_fragments.sh)
  2. Idealize protein structure (design_ideal.sh)
  3. Fixed back bond design (design_fix.sh)
  4. Flexible back bond design (design_flex.sh)

- **Major parameters that control flexible backbone protein design**
  1. Residue Definition File
  2. Extended rotamer libraries : ex1, ex2, ex3

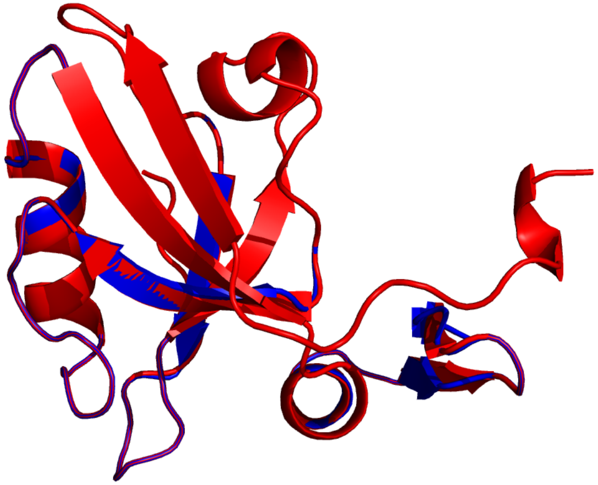
Rosetta Protein Design: An Example

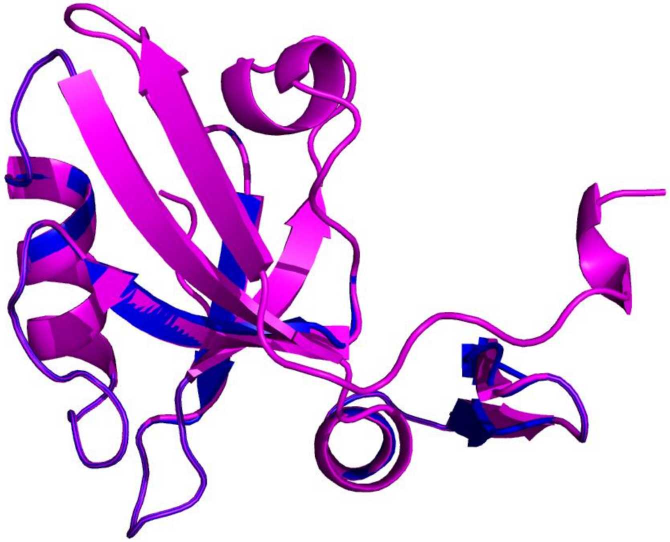
- **Command**
  Run_1be9.bat

- **Residue definition file**
  1be9.resfile

- **Flag file**
  Flag

- **Model selection**





Blue: X-Ray
Red: 1be9_0555
Score: -223.848

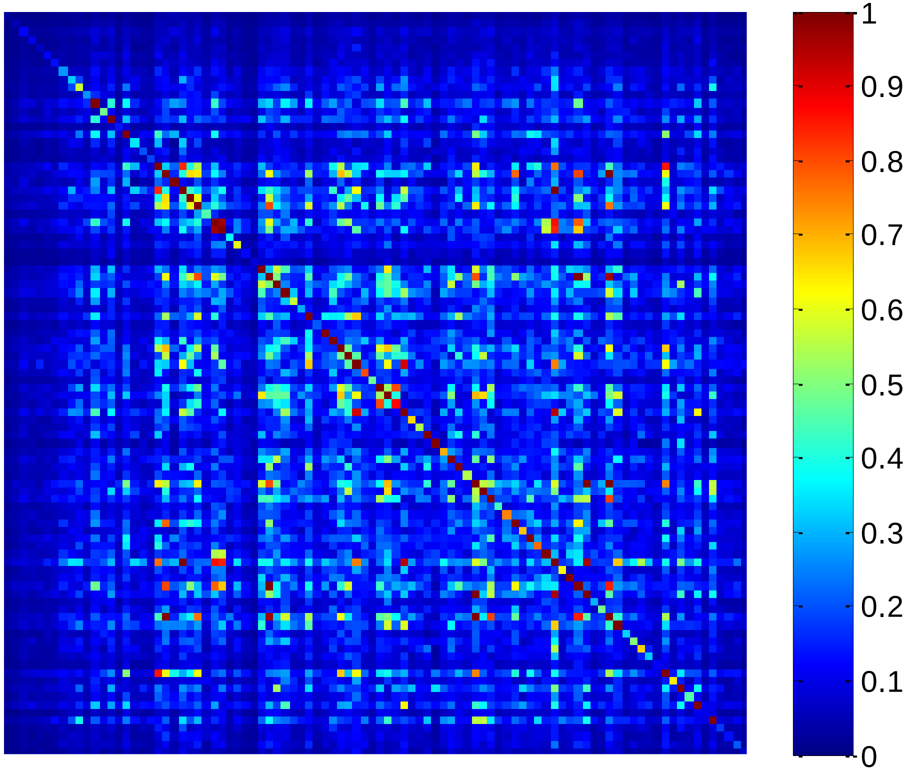
Magenta: 1be9_0278
Score: -215.518

Plot of CMR of 240 PDZ Sequences

1 11 21 31 41

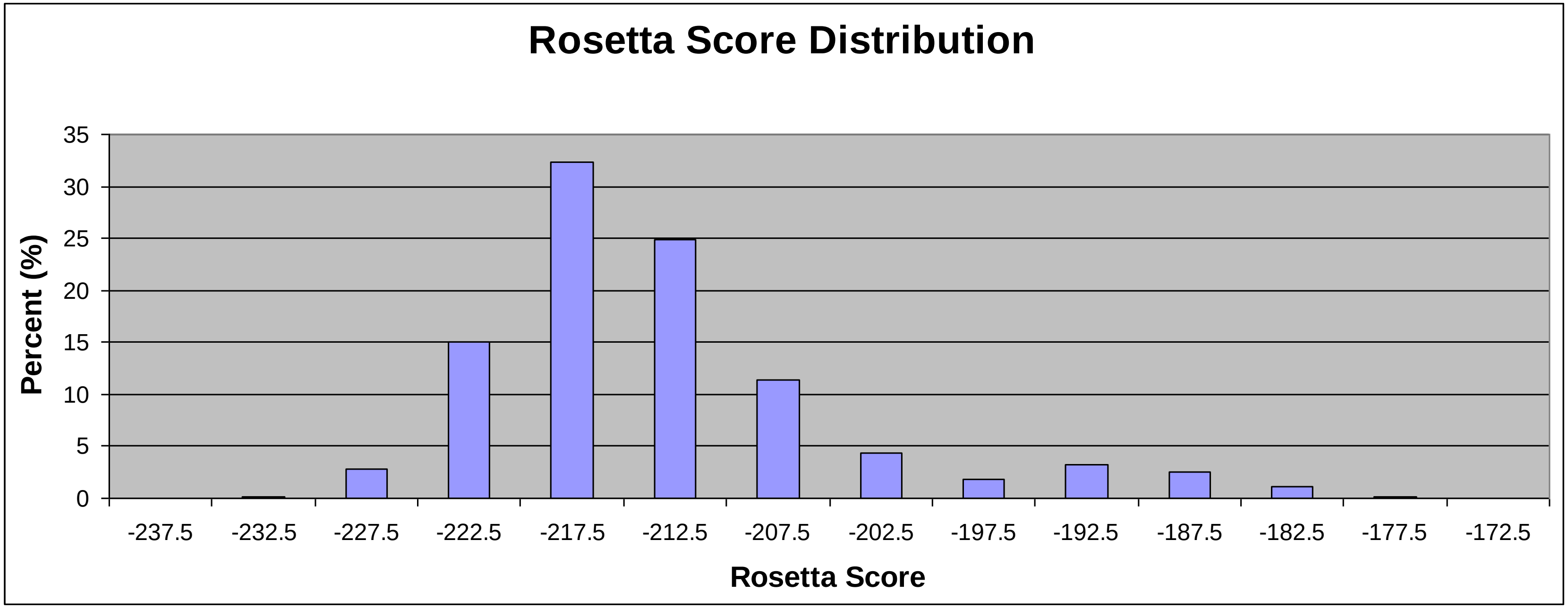
GEEDIPREPRRIVIHRGSTGLGFNIIGGED AGGPADLSGE

LRKGDQILSVNGVDLRNASHEQAAIALKNA KPEE

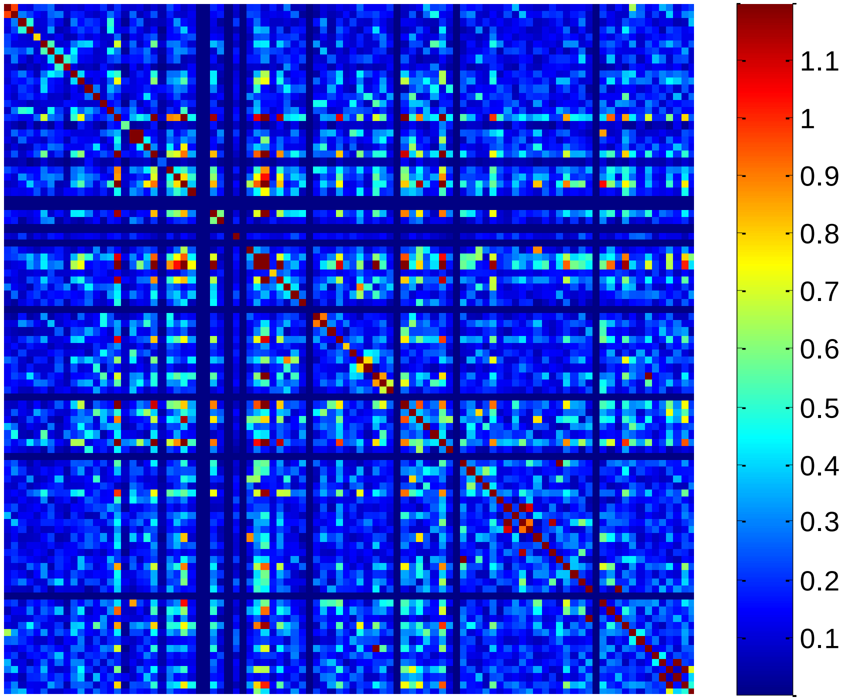


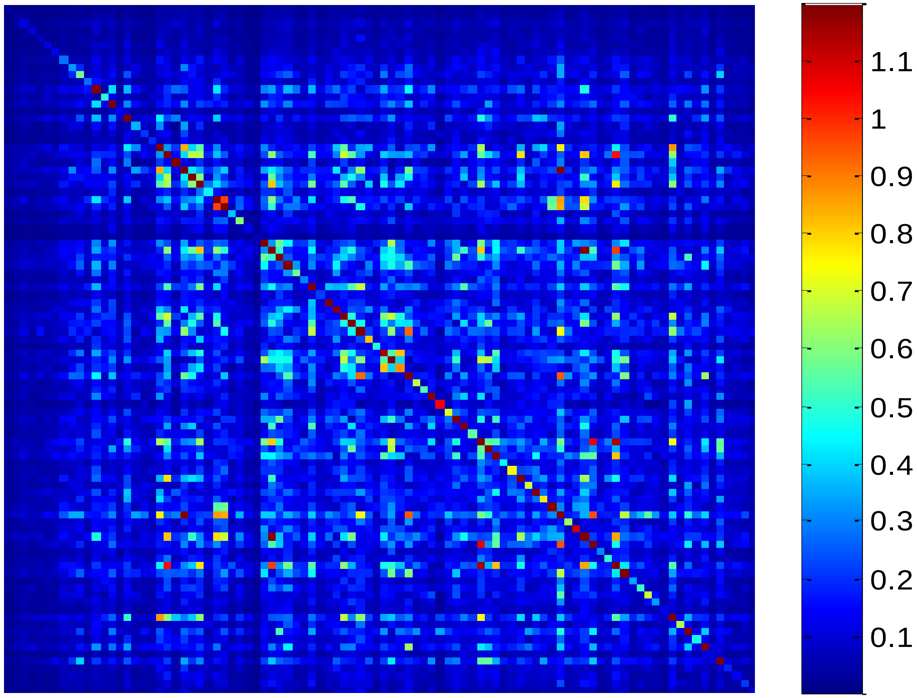
23-70
34-73
60-77
60-74
21-77
34-77
70-74
51-70
45-51
25-34

Distribution of Rosetta Scores



Statistical Coupling Analysis of 240 Rosetta Sequences

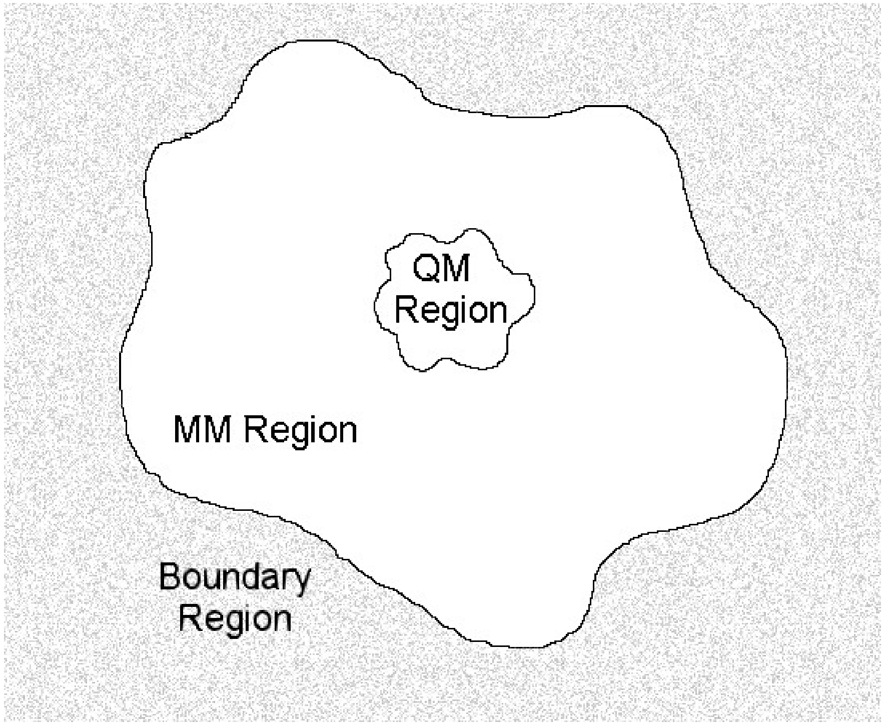




Modeling Reaction

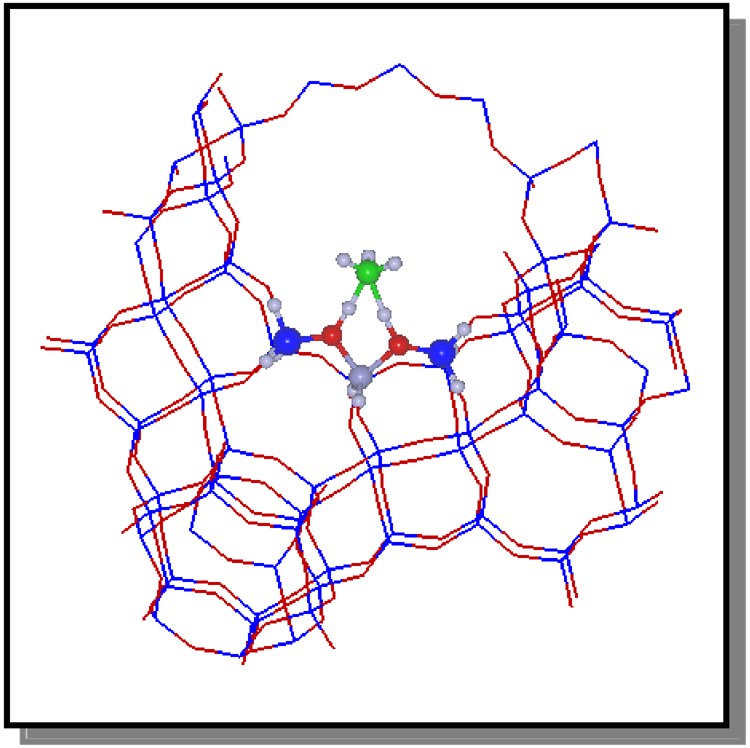
A Hybrid QM/MM Approach

The development of hybrid QM/MM approaches is guided by the general idea that large chemical systems may be partitioned into an electronically important region which requires a quantum chemical treatment and a remainder which only acts in a perturbative fashion and thus admits a classical description.



The QM/MM Modelling Approach

- Couple quantum mechanics and molecular mechanics approaches
- QM treatment of the active site
  - reacting centre
  - excited state processes (e.g. spectroscopy)
  - problem structures (e.g. complex transition metal centre)
- Classical MM treatment of environment
  - enzyme structure
  - zeolite framework
  - explicit solvent molecules
  - bulky organometallic ligands



QM/MM Methods

- Construct a Hamiltonian for the system consisting of a QM region and an MM region

$$H = H_{QM} + H_{MM} + H_{QM/MM}$$

- QM and MM regions interact mechanically and electronically (electrostatics, polarization)

- If bonds cross boundary between QM and MM region:
  - Cap bonds of QM region with link atoms
  - Use frozen or hybrid orbitals to terminate QM bonds

The Simplest Hybrid QM/MM Model

Hamiltonian for the molecular system in the Born-Oppenheimer approximation:

$$H = -\frac{1}{2}\sum_{i}^{electrons}\nabla^2 - \sum_{i}^{electrons}\sum_{j}^{nuclei}\frac{Z_j}{R_{ij}} + \sum_{i}^{electrons}\sum_{j<i}^{electrons}\frac{1}{r_{ij}} + \sum_{i}^{nuclei}\sum_{j<i}^{nuclei}\frac{Z_i Z_j}{R_{ij}}$$

←"Standard" QM Hamiltonian

$$H = -\frac{1}{2}\sum_{i}^{electrons}\nabla^2 - \sum_{i}^{electrons}\sum_{j}^{nuclei}\frac{Z_j}{R_{ij}} + \sum_{i}^{electrons}\sum_{j<i}^{electrons}\frac{1}{r_{ij}} + \sum_{i}^{nuclei}\sum_{j<i}^{nuclei}\frac{Z_i Z_j}{R_{ij}} \underbrace{- \sum_{i}^{electrons}\sum_{k}^{charges}\frac{Q_k}{R_{ik}} + \sum_{i}^{nuclei}\sum_{k}^{charges}\frac{Z_i Q_k}{R_{ik}}}_{Effect\ of\ External\ Charges}$$

The main drawbacks of this simple QM/MM model are:

- it is impossible to optimize the position of the QM part relative to the external charges because QM nuclei will collapse on the negatively charged external charges.
- some MM atoms possess no charge and so would be invisible to the QM atoms
- the van der Waals terms on the MM atoms often provide the only difference in the interactions of one atom type versus another, i.e. chloride and bromide ions both have unit negative charge and only differ in their van der Waals terms.

A Hybrid QM/MM Model

So, it is quite reasonable to attribute the van der Waals parameters (as it is in the MM method) to every QM atom and the Hamiltonian describing the interaction between the QM and MM atoms can have a form:

$$\hat{H}_{QM/MM} = -\sum_{i}^{electrons}\sum_{j}^{MM\,atoms}\frac{Q_j}{r_{ij}} + \sum_{i}^{nuclei}\sum_{j}^{MM\,atoms}\frac{Z_i Q_j}{R_{ij}} + \sum_{i}^{nuclei}\sum_{j}^{MM\,atoms}\left\{\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right\}$$

The van der Waals term models also electronic repulsion and dispersion interactions, which do not exist between QM and MM atoms because MM atoms possess no explicit electrons.

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(1976), 227-49***

The Hybrid QM/MM Model

Now we can construct a "real" hybrid QM/MM Hamiltonian:

$$\hat{H} = \hat{H}_{QM} + \hat{H}_{QM/MM} + \hat{H}_{MM}$$

$$\boxed{\hat{H}_{QM/MM} = -\sum_{i}^{electrons}\sum_{j}^{MM\,atoms}\frac{Q_j}{r_{ij}} + \sum_{i}^{nuclei}\sum_{j}^{MM\,atoms}\frac{Z_iQ_j}{R_{ij}} + \sum_{i}^{nuclei}\sum_{j}^{MM\,atoms}\left\{\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right\}}$$

$$\hat{H}_{QM} = -\frac{1}{2}\sum_{i}^{electrons}\nabla^2 - \sum_{i}^{electrons}\sum_{j}^{nuclei}\frac{Z_j}{R_{ij}} + \sum_{i}^{electrons}\sum_{j<i}^{electrons}\frac{1}{r_{ij}} + \sum_{i}^{nuclei}\sum_{j<i}^{nuclei}\frac{Z_iZ_j}{R_{ij}}$$

$$\hat{H}_{MM} = \sum_{bonds}K_b(R-R_0)^2 + \sum_{angles}K_\theta(\theta-\theta_0) + \sum_{dihedrals}\frac{V_\varphi}{2}(1+\cos(n\varphi)) + \sum_{nonbonded}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_iq_j}{R_{ij}}\right]$$

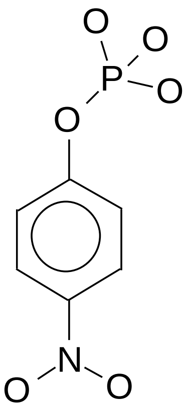
A "standard" MM force field can be used to determine the MM energy. For example, AMBER-like force field has a form:

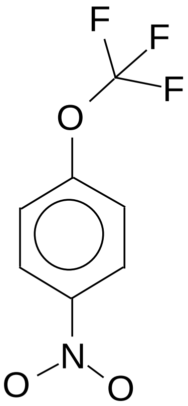
Choice of QM method

... is a compromise between computational efficiency and practicality and the desired chemical accuracy.

The main advantage of semi-empirical QM methods is that their computational efficiency is orders of magnitude greater than either the density functional or ab initio methods

Calculation times (in time units)





1800	RHF/6-31G*	36228
1	PM3	1

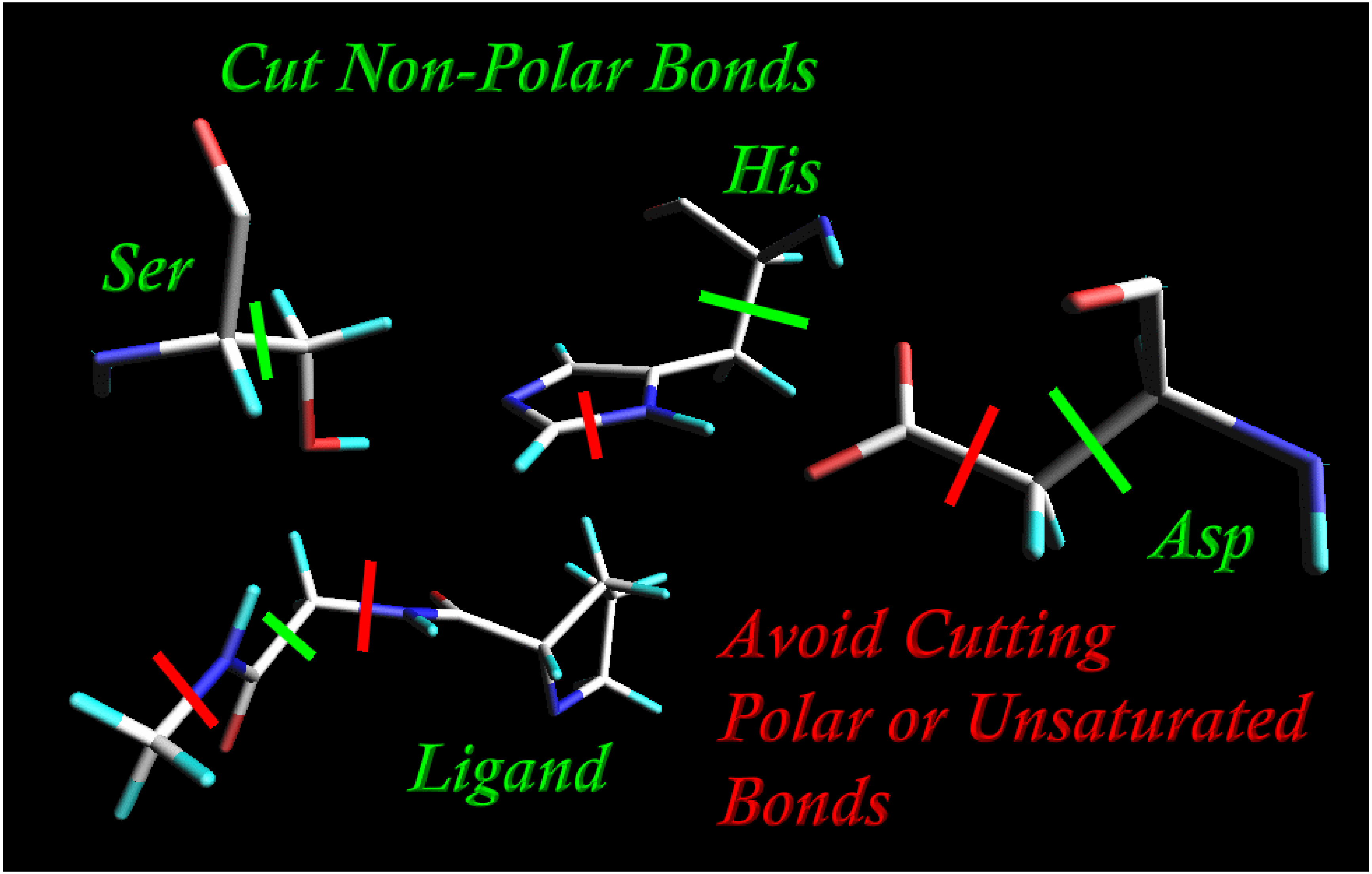
Hints for running QM/MM calculations

Choosing the QM region

- There are no good universal rules here
- One might want to have as large a QM region as possible
- However, having more than 80-100 atoms in the QM region will lead to simulations that are very expensive.
- for many features of conformational analysis, a good MM force field may be better than a semi-empirical or DFTB quantum description.

Hints for running QM/MM calculations

Choosing the QM region



Hints for running QM/MM calculations

Parallel Simulations

- At present all parts of the QM simulation are parallel except the density matrix build and the matrix diagonalisation.

- For small QM systems these two operations do not take a large percentage of time and so acceptable scaling can be seen to around 8 cpus.

- However, for large QM systems the matrix diagonalization time will dominate and so the scaling will not be as good.

Two Scenarios of Hybrid QM/MM

- Boundary through space (solute (QM) + solvent (MM))
  1. Unpolarized interaction: Solute (QM) + Solvent (MM)
  2. Polarized QM/unpolarized MM
  3. Fully polarized interactions

- Boundary through bond
  Link atoms

Empirical valence bond: a method related to QM/MM

- EVB attempts to combine empirical potential energy functions with valence bond ideas to describe chemical reactions efficiently and accurately.

- EVB starts with a N×N potential energy matrix:
  N diabatic states (diagonal)
  N(*N*-1) couplings (off-diagonal)

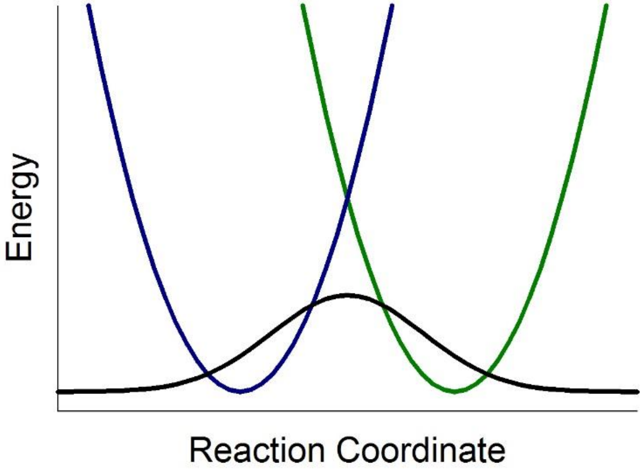
- Each diabatic state looks like a configuration in a standard non-reactive force field.

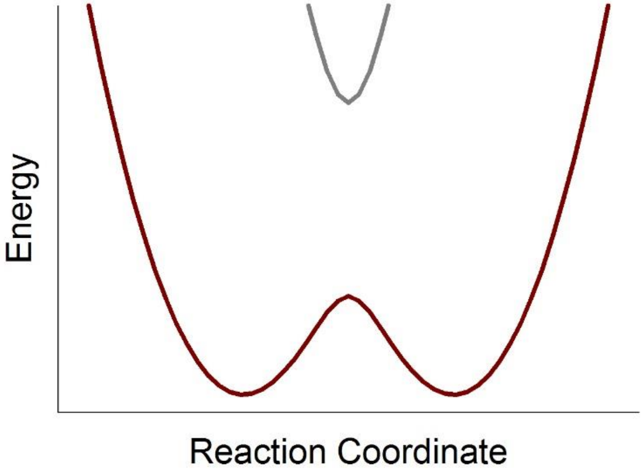
- Off-diagonal coupling elements: interaction between each diabatic state and the N-1 remaining states.

- Diagonalize V → adiabatic states. The minimal value is the ground state.

$$\phi_1 = HO^a - H^* + O^b H_2$$

$$\phi_2 = HO^{a-} + H^* + O^b H_2^+$$





Basic Idea - to be continued

If two diabatic states …

$$V = \begin{vmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{vmatrix}$$

$$\begin{vmatrix} V_{11} - E & V_{12} - ES_{12} \\ V_{21} - ES_{21} & V_{22} - E \end{vmatrix} = 0$$

Secular Equation

$$V = \frac{1}{2}(V_{11} + V_{22}) \pm \left[\left(\frac{V_{11} - V_{22}}{2} \right)^2 + V_{12}^2 \right]^{1/2}$$

Overlap integral is neglected

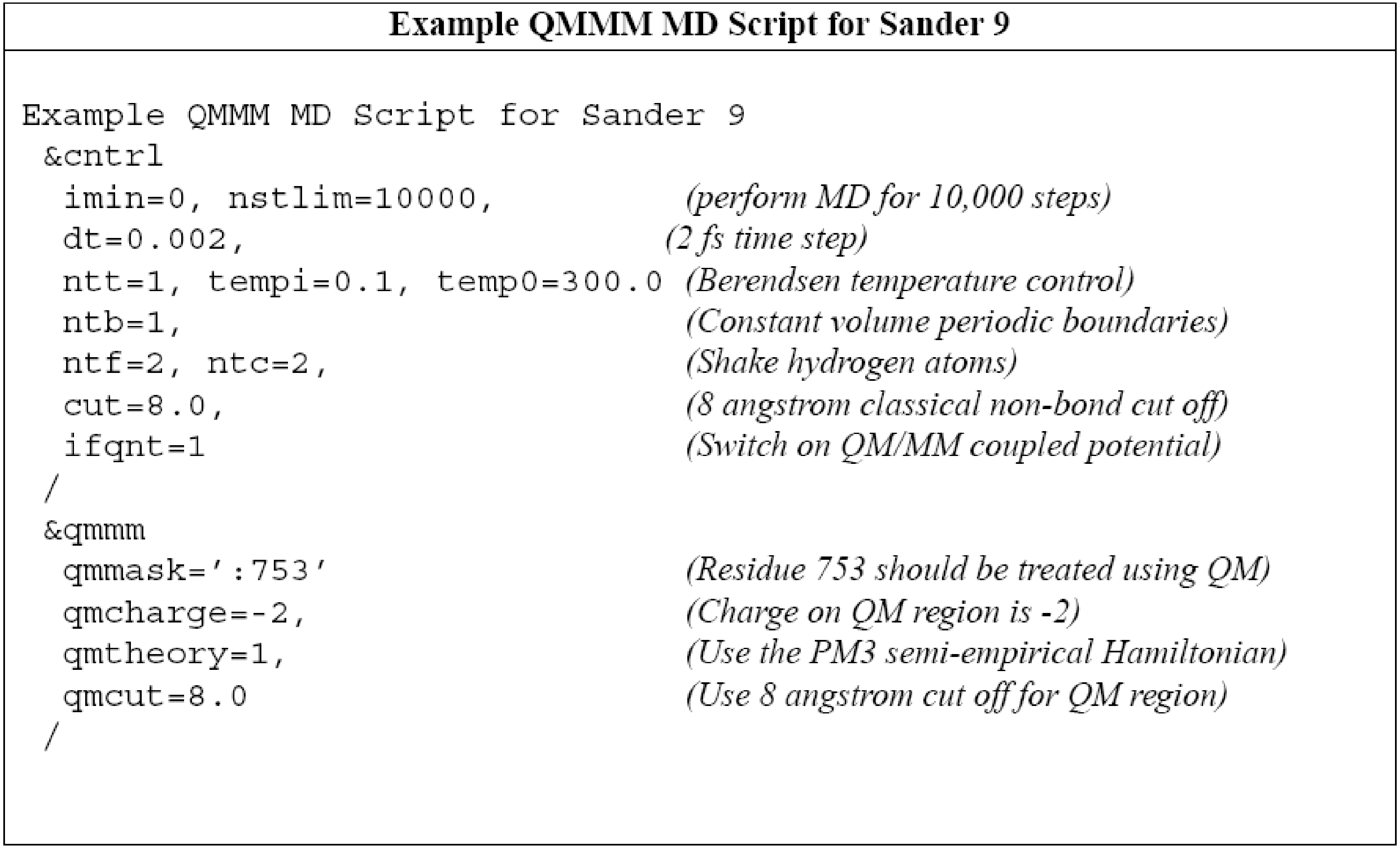
Amber QM/MM

- Amber features new and significantly improved QM/MM support

- The QM/MM facility supports gas phase, implicit solvent (GB) and periodic boundary (PME) simulations

- Compared to earlier versions, the QM/MM implementation offers improved accuracy, energy conservation, and performance.

Amber QM/MM Example



Example QMMM MD Script for Sander 9

```
Example QMMM MD Script for Sander 9
 &cntrl
  imin=0, nstlim=10000,                (perform MD for 10,000 steps)
  dt=0.002,                           (2 fs time step)
  ntt=1, tempi=0.1, temp0=300.0       (Berendsen temperature control)
  ntb=1,                               (Constant volume periodic boundaries)
  ntf=2, ntc=2,                        (Shake hydrogen atoms)
  cut=8.0,                             (8 angstrom classical non-bond cut off)
  ifqnt=1                              (Switch on QM/MM coupled potential)
 /
 &qmmm
  qmmask=':753'                        (Residue 753 should be treated using QM)
  qmcharge=-2,                         (Charge on QM region is -2)
  qmtheory=1,                          (Use the PM3 semi-empirical Hamiltonian)
  qmcut=8.0                            (Use 8 angstrom cut off for QM region)
 /
```

Amber QM/MM Example

Sample output

Summary of MMMS Couse

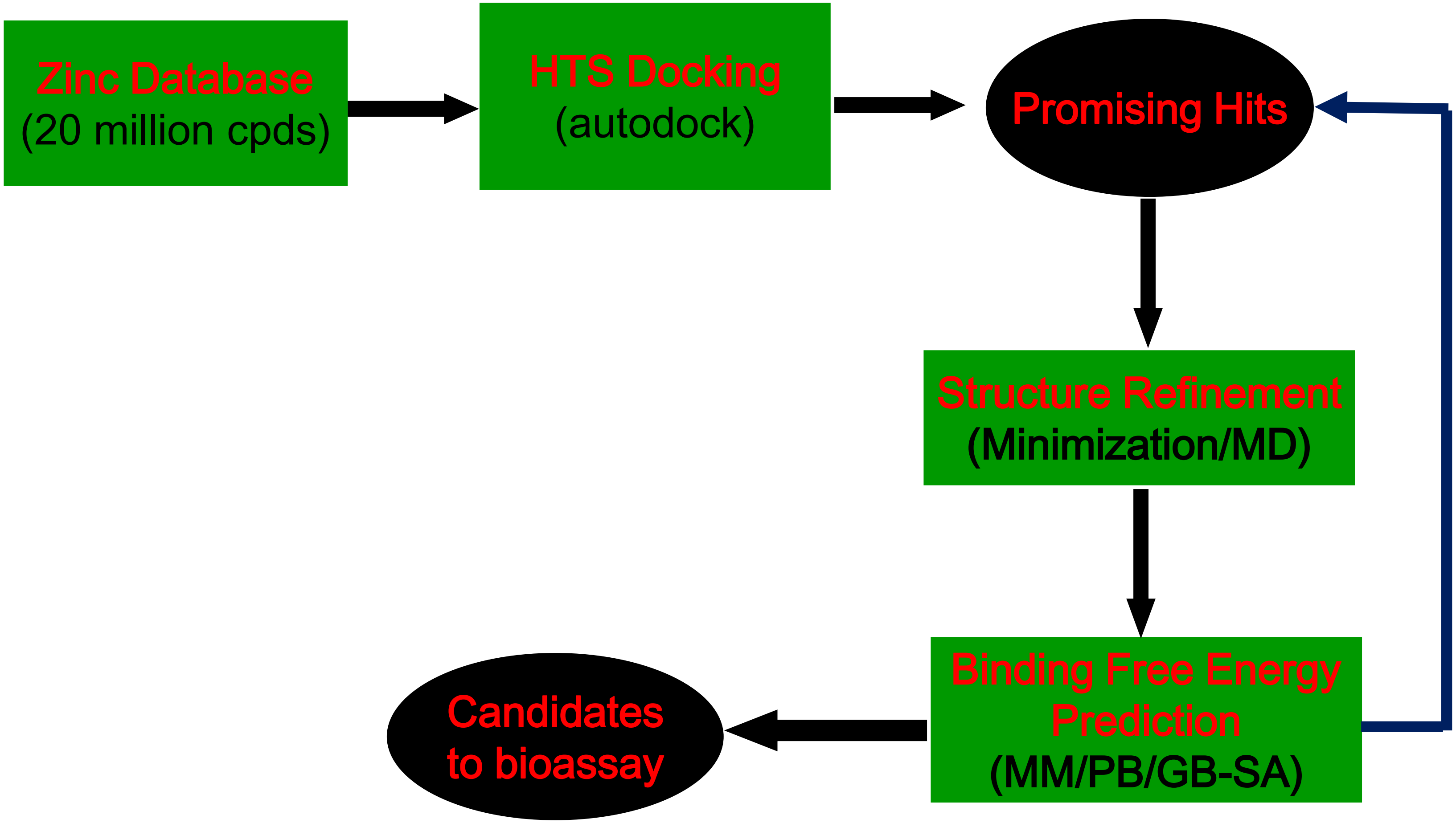
MM&MS Curriculum:

https://Mulan.swmed.edu/mmms/molecular_modelling.html

Application of MMMS in Biomedical Research

- Structure refinement
- Protein function
  1. Mutagenesis
  2. Dynamics
- Drug design

Basic Approaches of Free Energy Calculations



Zinc Database
(20 million cpds)

HTS Docking
(autodock)

Promising Hits

Structure Refinement
(Minimization/MD)

Binding Free Energy Prediction
(MM/PB/GB-SA)

Candidates to bioassay

Thank You For Your Attention

- Ajax Accounts (to be active until the end of 2016)
- Project Summary/Slides (in 4 weeks)
- TACC
- Computer for modeling