Protein Structure Prediction and Refinement

Chaok Seok
Department of Chemistry
Seoul National University

Structural and biophysical methods for biological macromolecules in solution
EMBO global exchange lecture course, Sungkyunkwan University, Jun 19-26, 2016
Contents

- Introduction: principles of structure prediction methods
- Computational prediction methods: theory and practice
  - Protein structure prediction (Robetta, 1-TASSER, HHpred, SWISS-MODEL)
  - Protein structure refinement, loop modelling, and docking (GALAXY)
- Conclusion: when/how to use computational methods
Introduction: principles of structure prediction methods

Natural law

Principles of physics or physical chemistry

Evolutionary principle

Related sequences, structures, functions
Physical principles for structure prediction

Experimentally observed states (or structures) corresponds to thermodynamically stable states in the given condition (pH, salt concentration, temperature, etc).

Protein molecules often have enough time to reach stable states before they act $\rightarrow$ reproducible/reliable function (in equilibrium)

Physical law for stability?

Stable states have the lowest free energy

$G=H-TS$

Transitions between states corresponds to the minimum-free energy path
Free energy as a function of state (structure/conformation)

\[ G \text{ (Free energy at constant pressure)} = H \text{ (enthalpy)} - T \text{ (temperature)} S \text{ (entropy)} \]

\[ H \text{ (enthalpy)} = E \text{ (energy)} - pV \text{ (energy needed to maintain pressure)} \]

\[ E = \text{ potential energy (interaction energy)} + \text{ kinetic energy (determined by temperature)} \]

Physical principle for potential energy of molecules

Quantum mechanics

Energy is determined by electrostatic interactions between electrons and nuclei
Quantum Mechanics of Many-Electron Systems (Dirac ’29)

“The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations that are much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to explanation of the main features of complex atomic systems without too much computation.”
Nobel prize in chemistry, 1998

Chemistry with computers

Chemistry is not only test tubes and chemicals. In quantum chemistry, quantum mechanics is used to compute the properties of molecules and their interactions. The Nobel laureates have made it possible to use quantum mechanics to study molecules with the help of computers.

\[ H \Psi = E \Psi \]

Charge distributions

Charge distributions in molecules can be computed with quantum-chemical methods. Excess of electrons (red in the illustration) in one molecule is drawn to a more positively charged portion of another molecule (blue). In this way the base pairs bind DNA together.
Computer time required for quantum calculations

Example: water molecules

1 water molecule: \(~1\) min

10 water molecules: \(~1\) day (for a single configuration)

10 water molecules, 1 microsecond: \(>1000\) days

10000 water molecules, 1 microsecond: \(>\)million years

\(-\rightarrow\) approximate energy with that of molecular mechanics (classical mechanics)
Molecular mechanics energy

\[ U = \sum \frac{1}{2} K_b (b-b_0)^2 + \sum \frac{1}{2} K_\theta (\theta-\theta_0)^2 + \sum K_\phi [1 - \cos(n\phi + \delta)] + \sum \varepsilon \left[ \left( \frac{r}{r_0} \right)^n - 2 \left( \frac{r}{r_0} \right)^6 \right] + \sum \frac{332q_i q_j}{r} \]

All Hooke
All Torsion
Fourier
All Nonbonded
Van der Waals
All partial
Coulomb

Simple sum over many terms
The Nobel Prize in Chemistry 2013
Martin Karplus, Michael Levitt, Arieh Warshel

The Nobel Prize in Chemistry 2013

“Development of Multiscale Models for Complex Chemical Systems”
The Nobel Prize in Chemistry 2013

Warshel & Levitt 1976
QM/MM: To study enzymatic reactions, we divide the system in two parts (Warshel & Levitt, JMB 1976)
States on the energy surface

**MOVING OVER ENERGY SURFACE**

- EM: Energy Minimization drops into local minimum.
- NMD: Normal Mode Dynamics vibrates about minimum.
- MD: Molecular Dynamics uses thermal energy to move smoothly over surface.
“...everything that living things do can be understood in terms of the jiggling and wiggling of atoms.”

The Feynman Lectures in 1963
States observed by molecular dynamics simulation

Correspond to experimentally observed states in the same condition (pH, salt, temperature, etc) if

the potential energy function is accurate,

and

the simulation is long enough (in the experimental time scale).

Problems of force field:

Simple functional form, Parameter training on small molecules, + Heuristic adjustments
Lack of polarization,
First molecular dynamics (MD) simulation of protein BPTI Simulation (9.2ps)
Computer time required for a folding protein by molecular dynamics simulation

A millisecond-simulation of a protein in solution with a time step of 1 femtosecond requires $10^{12}$ steps!

(each step requires energy and gradient evaluation for protein and all solvent molecules)

$\rightarrow$ new method needed!
Failure of Force Field in Protein Structure Prediction

David E. Shaw

Anton

Energy problem
CASP (Critical Assessment of techniques for protein Structure Prediction)

Community-wide experiment on the comparative evaluation of protein structure prediction methods

held every two years since 1994
Refinement of protein structure homology models via long, all-atom molecular dynamics simulations

Closerness from the crystal structure

David E. Shaw
Knowledge-based Potential in Protein Structure Prediction

Inverse-Boltzmann law

\[ E(r_i, r_j) = -kT \log \frac{N(r_{ij})}{N_{\text{ref}}(r_{ij})} \]

N from the structure database

A popular type of free energy function used in structure prediction.
Stable states in the free energy surface

MD simulation
potential energy -> sampling -> distribution & free energy

Global optimisation
global minimum (& sometimes suboptimal minima), but not (thermodynamic) distribution
However, ab initio structure prediction is still extremely difficult.

Reports on successful predictions for small proteins (~100aa)

Best predictors in CASP: Rosetta (Robetta), 1-TASSER

Recent advance (by Rosetta) extract contact information from related sequences (GREMLIN) -> restraint optimisation

![eLife](https://eLife.com)
Protein structure prediction

Ab initio prediction is usually inaccurate.

Homology modeling
(or Template-based modeling based on evolutionary principle or Comparative modeling)

can be very accurate if close template(s) can be found.
Homology modeling

Ab initio prediction is usually inaccurate.

Homology modeling
(or Template-based modeling based on evolutionary principle
or Comparative modeling)

can be very accurate if close template(s) can be found.
Procedure of homology modeling

- Domain parsing
  - Template selection
  - (Multiple) sequence alignment
  - Model building
  - Loop modeling, Refinement
Web servers for protein structure prediction

- I-TASSER
  - http://zhanglab.ccmb.med.umich.edu/I-TASSER/
- Robetta
  - http://robetta.bakerlab.org/
- HHpred
  - https://toolkit.tuebingen.mpg.de/hhpred
- SWISS-MODEL
Example: IL-4

- Interleukine-4 in Sus scrofa (Pig)
- Involved in at least several B-cell activation processes as well as in other cell types.
- Its structure has not been determined yet, but structures of homologs in other organisms have been determined.

http://www.uniprot.org/uniprot/F1CFE1
I-TASSER

- One of the best prediction servers in recent CASPs
- Automatic prediction server without any intermediate steps for the user
- Requires registration
- Predicts:
  - Protein tertiary structure
  - Ligand binding sites
  - Functions (GO, ...)
I-TASSER

http://zhanglab.ccmb.med.umich.edu/I-TASSER/
I-TASSER

I-TASSER On-line Server (View an example of I-TASSER output):

Copy and paste your sequence below ([10, 1500] residues in FASTA format). Click here for a sample input:

>tr/F1CFE1/F1CFE1 Pig interleukin-4 OS=Sus scrofa GN=IL-4 PE=2 SV=1
MGLTSQLIPTVCLLACTSNFVHGKCDTALOEIKTLNILTARKNCSMELPVDVFAAP
ENTTEKETFRCASTVLRHIYRHHTCMKSSLGLDRNLSSMANMTCSVHEAKKSTKDFLE
RLKTIMKEKYSKC

Or upload the sequence from your local computer:

Email: (mandatory, where results will be sent to)
chaok@snu.ac.kr

Password: (mandatory, please click here if you do not have a password)
******

ID: (optional, your given name of the protein)
IL-4

Option I: Assign additional restraints & templates to guide I-TASSER modeling.

Option II: Exclude some templates from I-TASSER template library.

Option III: Specify secondary structure for specific residues.

Keep my results public (uncheck this box if you want to keep your job private. A key will be assigned for you to access the results)

Run I-TASSER Clear form
(Please submit a new job only after your old job is completed)
The sequence has been successfully submitted to the I-TASSER server.

Your job id number is S279887. You will be notified by email once the job is completed if an email has been provided. The results of structure and function predictions for your submitted sequence will be available at (You may bookmark this link for your convenience of future visit): http://zhanglab.ccmb.med.umich.edu/I-TASSER/output/S279887/.

Your submitted sequence is of 133 residues:
>IL-4
MGLTSQLIPTLVCLACTSNFVHGHKCDITLQEIIKTLNILTARKNSCMELPVTDVFAAP
ENTTEKETFCRASTVLRHIYRHHTCMKSSLGDLRNLSSMANMTCSVHEAKKSTLKDFLE
RLKTIMKEKYSKC

Bookmark this link or
The server will sent an e-mail when the prediction is done.

It takes ~72 hours for a single prediction.
I-TASSER

- On the “Queue” page, you can track the job status.
I-TASSER

- Secondary Structure / Solvent Accessibility Prediction
I-TASSER

- Predicted normalized B-factor, which is related to per-residue model accuracy
I-TASSER

Top 5 final models predicted by I-TASSER

(For each target, I-TASSER simulations generate a large ensemble of structural conformations, called decoys. To select the final models, I-TASSER uses the SPICKER program to cluster all the decoys based on the pair-wise structure similarity, and reports up to five models which corresponds to the five largest structure clusters. The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5, 2], where a C-score of a higher value signifies a model with a higher confidence and vice-versa. TM-score and RMSD are estimated based on C-score and protein length following the correlation observed between these qualities. Since the top 5 models are ranked by the cluster size, it is possible that the lower-rank models have a higher C-score in rare cases. Although the first model has a better quality in most cases, it is also possible that the lower-rank models have a better quality than the higher-rank models as seen in our benchmark tests. If the I-TASSER simulations converge, it is possible to have less than 5 clusters generated; this is usually an indication that the models have a good quality because of the converged simulations.)

- More about C-score
- Local structure accuracy profile of the top five models

(By right-click on the images, you can export image file or change the configurations, e.g. modifying the background color or stopping the spin of your models)

- Download Model 1
  - C-score=0.13 (Read more about C-score)
  - Estimated TM-score = 0.73±0.11
  - Estimated RMSD = 3.9±2.7Å

- Download Model 2
  - C-score = -1.07

- Download Model 3
  - C-score = -1.08

It provides up to 5 models

It provides estimated model accuracies
I-TASSER

It predicts ligand binding sites

Predicted function using COACH

(This section reports biological annotations of the target protein by COACH based on the I-TASSER structure prediction. COACH is a meta-server approach that combines multiple function annotation from the COFACTOR, TM-SITE and S-SITE programs.)

Ligand binding sites

<table>
<thead>
<tr>
<th>Click to view</th>
<th>Rank</th>
<th>C-score</th>
<th>Cluster size</th>
<th>PDB Hit</th>
<th>Lig Name</th>
<th>Download Complex</th>
<th>Ligand Binding Site Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.95</td>
<td>194</td>
<td>2a7pC</td>
<td>GSH</td>
<td>Rep, Mult</td>
<td>20,22,23,25,55,66,67,68,79,80,81</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.05</td>
<td>9</td>
<td>2yanB</td>
<td>GSH</td>
<td>Rep, Mult</td>
<td>20,22,23,25,55</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.04</td>
<td>5</td>
<td>3fzaA</td>
<td>RME</td>
<td>Rep, Mult</td>
<td>25,28,80,81,84</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.02</td>
<td>5</td>
<td>3rhgB</td>
<td>k-mer</td>
<td>N/A</td>
<td>22,23,24,25</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.02</td>
<td>4</td>
<td>3mszA</td>
<td>GSH</td>
<td>Rep, Mult</td>
<td>25,66,68,79,80,81,82</td>
</tr>
</tbody>
</table>

Download the residue-specific ligand binding probability, which is estimated by SVM.
Download the all possible binding ligands and detailed prediction summary.
Download the templates clustering results.

(a) C-score is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a better prediction.
(b) Cluster size is the total number of templates in a cluster.
(c) Lig Name is name of possible binding ligand. Click the name to view its information in the BioLIP database.
(d) Rep is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the Mult is the complex structures with all potential binding ligands in the cluster.
I-TASSER

It also predicts GO terms and their reliability
Robetta

- One of the best prediction servers in recent CASPs
- Requires registration
- Automatic prediction server with a single pause for the user
  - It predicts domains first.
  - And you have to re-submit a prediction job for each domain.
- You can run one prediction job at a time.
Robetta

http://robetta.bakerlab.org/
Submit a job to the Server

If you submit more than one job using different logins, the jobs will be deleted and the IP may be locked out.

Required

Prediction Type: Ginzu: Domain Prediction
Structure: 3-D Model (available per domain after Ginzu completes from results page)

Registered Username: chaok
or
Registered Email Address:

Target Name: IL-4

Paste Fasta (AA sequence only!)

Or Upload Fasta:

Optional

Reply Email:

Do not warn me if my sequence matches one already submitted.

Note: please do not submit known PDB sequences or CASP targets intentionally

Submit

It takes ~72 hours for a single prediction.
Robetta

- On the “Queue” page, you can track the job status.

<table>
<thead>
<tr>
<th>Num</th>
<th>ID</th>
<th>Status</th>
<th>Method</th>
<th>Username</th>
<th>Target</th>
<th>Target</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66648</td>
<td>Queued</td>
<td>Ginzu</td>
<td>bawer178</td>
<td>TodS</td>
<td>TodS</td>
<td>TodS</td>
</tr>
<tr>
<td>2</td>
<td>66647</td>
<td>Queued</td>
<td>Ginzu</td>
<td>meghau</td>
<td>H_doris_luciferase_4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>66646</td>
<td>Queued</td>
<td>Ginzu</td>
<td>b015mtech11003</td>
<td>tdp43-2c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>66645</td>
<td>Queued</td>
<td>Ginzu</td>
<td>bootpp</td>
<td>ctCIPP1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>66644</td>
<td>Queued</td>
<td>Ginzu</td>
<td>kellyj2</td>
<td>H_melpomene_luciferase_7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>66643</td>
<td>Queued</td>
<td>Ginzu</td>
<td>a.jahangiri</td>
<td>Omp34-no signal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>66642</td>
<td>Active</td>
<td>Ginzu</td>
<td>iphan</td>
<td>rickA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>66641</td>
<td>Queued</td>
<td>Ginzu</td>
<td>amishra2</td>
<td>1NYC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>66640</td>
<td>Queued</td>
<td>Ginzu</td>
<td>emte1992</td>
<td>hhhh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>66639</td>
<td>Active</td>
<td>Ginzu</td>
<td>divnand</td>
<td>ORF 66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>66638</td>
<td>Active</td>
<td>Ginzu</td>
<td>biotcorp</td>
<td>haru1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>66637</td>
<td>Complete</td>
<td>Ginzu</td>
<td>johnjames</td>
<td>set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>66636</td>
<td>Complete</td>
<td>Ginzu</td>
<td>jaboyre</td>
<td>m128a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>66635</td>
<td>Complete</td>
<td>Ginzu</td>
<td>efischer</td>
<td>kiaa1524</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>66634</td>
<td>Complete</td>
<td>Ginzu</td>
<td>ekaterinakarpova</td>
<td>NEP1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>66633</td>
<td>Complete</td>
<td>Ginzu</td>
<td>yuanjiaqil273</td>
<td>NCX1_Cyto1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>66629</td>
<td>Complete</td>
<td>Ginzu</td>
<td>anfa</td>
<td>IgE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>66628</td>
<td>Complete</td>
<td>Ginzu</td>
<td>jaobrien</td>
<td>PA2B3_BUNFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>66627</td>
<td>Complete</td>
<td>Ginzu</td>
<td>zhangxin</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>66626</td>
<td>Active</td>
<td>Structure</td>
<td>dkhago</td>
<td>Eveides_isabella_luciferase_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>66625</td>
<td>Active</td>
<td>Structure</td>
<td>rwmartin</td>
<td>D_plexippus_DPOGS206945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>66624</td>
<td>Complete</td>
<td>Ginzu</td>
<td>amakiriaustin</td>
<td>medin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>66623</td>
<td>Active</td>
<td>Ginzu</td>
<td><a href="mailto:renata.tisi@unimib.it">renata.tisi@unimib.it</a></td>
<td>ScRad50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>66622</td>
<td>Complete</td>
<td>Ginzu</td>
<td>sprajap</td>
<td>LD3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>66621</td>
<td>Queued</td>
<td>Structure</td>
<td>itoni</td>
<td>hc_full</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Status**
- Queued: waiting for running
- Active: running now
- Complete: finished

**Method**
- Ginzu: domain prediction stage

**Structure**
- tertiary structure prediction stage
Once you got "domain prediction" results, you can choose domains to predict 3D structures.

Press this button to predict as a “single” domain with “comparative modeling” method.
Robetta

- Once you got “domain prediction” results, you can choose domains to predict 3D structures.

It is predicted as 3-domain protein. You can model each domain with suggested modeling method (either TBM or ab initio).
Robetta

It provides up to 5 models and related predicted features.
HHpred

- One of the best prediction servers in recent CASPs that are the fastest.

- It does NOT require registration,
  but requires a MODELLER license key to build 3D model.

- It provides an interactive prediction mode
  and also provides automatic selection options.
HHpred

http://toolkit.tuebingen.mpg.de/hhpred
HHpred

Paste amino acid sequence

Check this for the better sequence alignment
HHpred

This bar shows aligned regions for each templates (red: reliable matches; blue: unreliable matches)

Press this menu for 3D model building

Detailed template detection informations
1. Manual template selection: Generate a PIR-alignment of your sequence with the selected template or templates in order to build a 3D model with MODELLER.

2. Automatic template selection: Optimize diversities of query and template HMMs, rerank templates and automatically select best set. In further steps a multiple alignment is created from this set, and a 3D model is build with MODELLER using this alignment.

Only hits found in PDB or SCOP can be used to create a model (other hits are disabled).

**HHpred**

**Manual template selection**

**Automatic template selection**
**HHpred**

HHpred makemodel - Results

Select templates

- optimal single template
- optimal multiple templates
- user-defined

You can build your model using an optimal single template per query sequence domain (HHpred4 in CASP8), optimal multiple templates (HHpred5 in CASP8) or your own selected templates. We recommend not to select more than 5 template matches per query sequence domain.

The hit list below may contain several matches per template with similar alignments. These differ by the diversity (thickness) of the query and template HMMs used to generate the HMM-HMM alignment. A neural network is used to predict the expected TMScore for each alignment (column 6) and template alignments are reranked by this score. (TMScore ∈ [0,1]; TMScore ≥ 0.4 corresponds to meaningful predictions.)

Press "Generate alignment for MODELLER" to generate an PIR-alignment from the (automatically or manually) selected alignments.

Press this button to generate sequence alignment between query and selected templates.
**HHpred**

You can see the model structure on web browser

You can predict model accuracies

<table>
<thead>
<tr>
<th>ATOM</th>
<th>CHAIN</th>
<th>RES</th>
<th>NAME</th>
<th>ALT</th>
<th>SERIAL</th>
<th>ECOD</th>
<th>QUALITY by VERIFY3D</th>
<th>QUALITY by SOLVEX</th>
<th>QUALITY by ANOLEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>44.056</td>
<td>58.247</td>
<td>2.669</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>CA</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>43.210</td>
<td>58.884</td>
<td>3.702</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>CB</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>41.782</td>
<td>58.316</td>
<td>3.644</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>CG</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>40.784</td>
<td>59.076</td>
<td>4.519</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>SD</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>39.056</td>
<td>58.575</td>
<td>4.269</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>CE</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>38.932</td>
<td>59.312</td>
<td>2.614</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>43.780</td>
<td>58.633</td>
<td>5.056</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>O</td>
<td>MET</td>
<td></td>
<td></td>
<td>1</td>
<td>44.869</td>
<td>59.105</td>
<td>5.379</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>GLY</td>
<td>2</td>
<td></td>
<td>2</td>
<td>43.053</td>
<td>57.867</td>
<td>5.891</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>CA</td>
<td>GLY</td>
<td>2</td>
<td></td>
<td>2</td>
<td>43.543</td>
<td>57.600</td>
<td>7.209</td>
<td>1.00</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>GLY</td>
<td>2</td>
<td></td>
<td>2</td>
<td>44.739</td>
<td>56.722</td>
<td>7.064</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>O</td>
<td>GLY</td>
<td>2</td>
<td></td>
<td>2</td>
<td>44.849</td>
<td>55.951</td>
<td>6.112</td>
<td>1.00</td>
</tr>
<tr>
<td>13</td>
<td>N</td>
<td>LEU</td>
<td>3</td>
<td></td>
<td>3</td>
<td>45.674</td>
<td>56.821</td>
<td>8.028</td>
<td>1.00</td>
</tr>
</tbody>
</table>
SWISS-MODEL

- It does NOT require registration.
- It provides both interactive and automatic predictions.
- It uses only single templates for each prediction.
SWISS-MODEL

http://swissmodel.expasy.org/
Both modes take few minutes for a single prediction.
SWISS-MODEL

- Interactive prediction mode

You can select templates to build 3D models
Here are predicted model accuracies

<table>
<thead>
<tr>
<th>Oligo-State</th>
<th>Ligands</th>
<th>GMQE</th>
<th>QMEAN4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MONOMER</td>
<td>None</td>
<td>0.71</td>
<td>-3.40</td>
</tr>
</tbody>
</table>

**Model-Template Alignment**

<table>
<thead>
<tr>
<th>Template</th>
<th>Seq Identity</th>
<th>Coverage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1bcn.1A</td>
<td>66.07%</td>
<td></td>
<td>INTERLEUKIN-4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oligo-State</th>
<th>Ligands</th>
<th>GMQE</th>
<th>QMEAN4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MONOMER</td>
<td>None</td>
<td>0.74</td>
<td>-1.53</td>
</tr>
</tbody>
</table>

**Model-Template Alignment**

<table>
<thead>
<tr>
<th>Template</th>
<th>Seq Identity</th>
<th>Coverage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b8z.1A</td>
<td>68.81%</td>
<td></td>
<td>Interleukin-4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oligo-State</th>
<th>Ligands</th>
<th>GMQE</th>
<th>QMEAN4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MONOMER</td>
<td>None</td>
<td>0.75</td>
<td>-0.57</td>
</tr>
</tbody>
</table>

**Model-Template Alignment**

<table>
<thead>
<tr>
<th>Template</th>
<th>Seq Identity</th>
<th>Coverage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b91.1A</td>
<td>68.81%</td>
<td></td>
<td>Interleukin-4</td>
</tr>
</tbody>
</table>

It predicts a model with for each template.
SWISS-MODEL

It provides estimated model accuracies (global) and per-residue accuracies.

This figure shows local model accuracies (blue means reliable).
GALAXY programs for protein structure prediction, refinement, and docking
Structure Prediction

- Galaxy TBM
- Galaxy Dom
- Galaxy Loop
- Galaxy Refine
- Galaxy Site
- Galaxy Dock
- Galaxy 7TM

Protein Docking

- Galaxy Gem
- Galaxy PPDock
- Galaxy PepDock

Binding site prediction

- Galaxy Site

Protein-ligand docking

- Galaxy Dock

GPCR docking

- Galaxy 7TM

Domain prediction

- Galaxy Dom

Loop modeling

- Galaxy Loop

Template-based modeling

- Galaxy TBM

Protein structure refinement

- Galaxy Refine

Oligomer prediction

- Galaxy Gemini

Protein-protein docking

- Galaxy PPDock

Protein-peptide docking

- Galaxy PepDock

Ligand Docking
GalaxyWEB server
(http://galaxy.seoklab.org)
What are the most used programs?

Out of 6,936 runs after excluding the jobs submitted through SNU IPs (as of Jan 17, 2016)
Protein loop modeling

Sequence Alignment

<table>
<thead>
<tr>
<th>sel=0</th>
<th>208</th>
<th>Seq:1</th>
<th>Pos:2201195</th>
<th>[T0424]</th>
<th></th>
<th>290</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0424</td>
<td>SPVATLCSSTIDRCNIEKIDEEIDTDVRFRFSEVTLAQSHGRSGDSA---KHDLYWYKDPILTTLHRPKTVVVSADNLAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1wruA</td>
<td>AT--DELV-------LGEnllTlDFeedFDRFRFSEVT K----------SRRGIAAEDEVTYRPMIIIADSKITAKDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3cddA</td>
<td>KAGVSRTI-------LDNVKaARCRF20ROCRFSKFT K----------AAGADVTDAEIGRPLIIVNEET-TAEGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3cddB</td>
<td>KAGVSRTI-------LDNVKaARCRF20ROCRFSKFT K----------GCGAADVTDAEIGRPLIIVNEET+TAEQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3cddD</td>
<td>KAGVSRTI-------LDNVKaARCRF20ROCRFSKFT K----------DTAELPTVGGCAADVTDAEIGRPLIIVNEETTAEQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2p5zX</td>
<td>KG--LTLP-------LRTAVGQMTAYZ-Z---N--- scaff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Loop Modeling
Variability of protein loops

Sequence and structure variability among homologous proteins

Functional specificity
Loop closure: an important issue in loop modeling

To close the backbone loop correctly between fixed structures.
Example of loop sampling

- **Experimental structure**
- **Sampled structures**
GalaxyLoop protein loop modeling

Loop Closure Algorithm

Fragment Assembly

Energy Function

Global Optimization

GalaxyLoop: blind prediction example on model framework structure

TR712

GDT-HA: 80.2 -> 82.1
HB-bb: 82.3 -> 92.4
MP score: 2.20 -> 1.28 (Polar H only)
GalaxyLoop server: input page
GalaxyLoop server: output page
GalaxyRefine: Overall relaxation

Mild relaxation (w/ stronger initial-structure restraint)

Aggressive relaxation (w/ weaker restraint)
GalaxyRefine: A successful blind prediction example

TR681

GDT-HA: 56.9 -> 63.9
HB-bb: 82.3 -> 92.4
MP score: 2.63 -> 1.44 (Polar H only)
GalaxyRefine server: input page

GalaxyWEB
A web server for protein structure prediction, refinement, and related methods
Computational Biology Lab, Department of Chemistry, Seoul National University

GalaxyRefine

Model structures generated by protein structure prediction methods can be refined. No gaps are allowed in the middle of initial protein structure.

User Information

Job name

E-mail address (Optional)

Model Structure to be refined

PDB File
(≤1000 AA)

Protein Structure File (allowed file extensions: pdb, txt)

Submit

submit reset

Help

- Information
- PDB File: File Format
- E-mail: Average run time is 1-2h. If e-mail address is given, the server sends notifications automatically. If not, the user has to bookmark the report page.

Example

- PDB file: TR747.pdb
- Report: [View]

Software

- You can download a standalone version at here.
GalaxyRefine server: output page
Community-wide blind prediction experiments

CASP (Critical Assessment of techniques for protein Structure Prediction)
Comparative evaluation of protein structure prediction methods

CAPRI (Critical Assessment of PRediction of Interactions)
Comparative evaluation of methods for prediction of protein-protein interactions

CSAR (Community Structure-Activity Resource)
Comparative evaluation of methods for prediction of protein-ligand interactions

GPCR Dock
Comparative evaluation of methods for GPCR modeling and docking
CAPRI

CAPRI (Critical Assessment of PRediction of Interactions)

Communitywide experiment on the comparative evaluation of methods for prediction of protein-protein interactions held irregularly

We participated since Round 20, Jan 2010.
### Prediction of Protein-Protein Interactions by GALAXY in CAPRI Round 30

<table>
<thead>
<tr>
<th>Participant</th>
<th>Participated targets</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seok</td>
<td>25</td>
<td>15/14**</td>
</tr>
<tr>
<td>Huang</td>
<td>25</td>
<td>16/13**</td>
</tr>
<tr>
<td>Guerois</td>
<td>25</td>
<td>16/12**</td>
</tr>
<tr>
<td>Zou</td>
<td>25</td>
<td>14/11**</td>
</tr>
<tr>
<td>Shen</td>
<td>25</td>
<td>13/11**</td>
</tr>
<tr>
<td>Grudinin</td>
<td>24</td>
<td>11/10**</td>
</tr>
<tr>
<td>Weng</td>
<td>25</td>
<td>13/9**</td>
</tr>
<tr>
<td>Vakser</td>
<td>25</td>
<td>11/9**</td>
</tr>
<tr>
<td>Vajda/Kozakov</td>
<td>24</td>
<td>15/8**</td>
</tr>
<tr>
<td>Fernandez-Recio</td>
<td>25</td>
<td>11/8**</td>
</tr>
<tr>
<td>Lee</td>
<td>20</td>
<td>10/7**</td>
</tr>
<tr>
<td>Tomii</td>
<td>20</td>
<td>8/6**</td>
</tr>
<tr>
<td>Sali</td>
<td>12</td>
<td>6/4**</td>
</tr>
<tr>
<td>Negi</td>
<td>25</td>
<td>7/3**</td>
</tr>
<tr>
<td>Eisenstein</td>
<td>6</td>
<td>3**</td>
</tr>
<tr>
<td>Bates</td>
<td>25</td>
<td>7/2**</td>
</tr>
<tr>
<td>Kihara</td>
<td>23</td>
<td>7/2**</td>
</tr>
<tr>
<td>Zhou</td>
<td>25</td>
<td>4/2**</td>
</tr>
<tr>
<td>Tovchigrechko</td>
<td>12</td>
<td>3/1**</td>
</tr>
<tr>
<td>Ritchie</td>
<td>8</td>
<td>2/1**</td>
</tr>
<tr>
<td>Fernandez-Fuentes</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Xiao</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Gray</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gong</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Del Carpio</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Wade</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Haliloglu</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Prediction of homo-oligomer structure by GALAXY

Target Protein

Oligomer Template Search (GalaxyGemini) and Model Building (GalaxyTBM)

- HHsearch on oligomer database
- Template selection with interface weighted rescoring
- Oligomer model building

Global Refinement with Symmetry Restraint (GalaxyRefine)

- Restraint from initial structure
- Complex structure refinement

Final Model

Symmetric ULR modeling (GalaxyLoop)

- ULR detection
- ULR modeling
Impact of refinement

<table>
<thead>
<tr>
<th>T85</th>
<th>$F_{\text{nat}}$</th>
<th>$F_{\text{nonnat}}$</th>
<th>LRMSD</th>
<th>IRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>0.60</td>
<td>0.32</td>
<td>4.02</td>
<td>2.27</td>
</tr>
<tr>
<td>After</td>
<td>0.66</td>
<td>0.34</td>
<td>2.31</td>
<td>2.07</td>
</tr>
</tbody>
</table>
GalaxyGemini:

Prediction of protein oligomeric structure is performed from protein monomer. The input protein structure may be either an experimental structure or a model structure.

User Information:

Job name
E-mail address (Optional)

Protein for oligomeric structure prediction:

PDB File
Protein Structure File (allowed file extensions: pdb, txt)

Energy minimization

Submit:

submit reset

Help:

- Information
- PDB File: File Format
- E-mail: Average run time is 30min~1h. If e-mail address is given, the server sends notifications automatically. If not, the user has to bookmark the report page address.
- Energy minimization: Energy minimization is used to remove steric clashes at the oligomer interface.

Example:

- PDB file: Gemini.pdb
- Report: [View]
GalaxyGemini server: output page
GalaxyRefineComplex server: input page
GalaxyRefineComplex server: output page
GalaxySite server: input page

GalaxyWEB
A web server for protein structure prediction, refinement, and related methods
Computational Biology Lab, Department of Chemistry, Seoul National University

GalaxySite
Prediction of ligand binding site of a query protein is performed. Up to three ligands that are likely to bind to the protein and their predicted binding poses are provided.

User Information
Job name
E-mail address (Optional)

Query Protein Information
SEQUENCE
(≤500 AA)

or PDB file
(≤500 AA) Protein Structure File (allowed file extensions: pdb, txt)

Submit
submit reset

Help
- Information
- PDB File: file format
- E-mail: Average run time ranges from 2h (for a structure input) to 4h (for a sequence input). If e-mail address is provided, the server sends notifications automatically. If not, the user has to bookmark the report page.

Examples
- Prediction from a protein sequence
  - SEQ: Uniprot Q8CNK1
  - Report: [View]
- Prediction from a protein structure
  - PDB file: SITE.pdb
  - Report: [View]
# GalaxySite server: output page

## FMN_binding

### Ligands predicted to bind

<table>
<thead>
<tr>
<th>No</th>
<th>Ligand Name</th>
<th>Ligand Structure</th>
<th>Template for protein-ligand complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FMN</td>
<td><img src="image1.png" alt="Structure" /></td>
<td>3esv_A, 1tjv_A, 3Nh_4_A, 128v_y, 1ykg_A, 38v_R, 174p_A, 269c_Y, 234t_A, 1yc_o, 1000_A, 12o2_A, 1agq_A, 3chu_A, 22p_A, 1650_A, 1ykg_A, 3chu_A, 22p_A, 3n3_d, 1tjv_A</td>
</tr>
<tr>
<td>2</td>
<td>FAD</td>
<td><img src="image2.png" alt="Structure" /></td>
<td>2topo_A</td>
</tr>
<tr>
<td>3</td>
<td>BEN</td>
<td><img src="image3.png" alt="Structure" /></td>
<td>2lmm_A</td>
</tr>
</tbody>
</table>

### Predicted ligand-binding residues

<table>
<thead>
<tr>
<th>No</th>
<th>Ligand Name</th>
<th>Binding Residues</th>
<th>Interaction Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FMN</td>
<td>3esv 27o 28t 30l 31n 75a 79t 76y 81n 113l 114g 115n 118y 120t 121p 122n 144q 152k</td>
<td>LINK</td>
</tr>
<tr>
<td>2</td>
<td>FAD</td>
<td>2st 30t 30y 115n 118y 148d</td>
<td>LINK</td>
</tr>
<tr>
<td>3</td>
<td>BEN</td>
<td>3gs 40k 42a 43h 45r 49g</td>
<td>LINK</td>
</tr>
</tbody>
</table>

### Predicted binding poses

![3D Model](image4.png)
GalaxyDom server: input page
## Modeling Unit Information

<table>
<thead>
<tr>
<th>No</th>
<th>Regions (Length)</th>
<th>Type</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-26 (26)</td>
<td>SEQADV</td>
<td>QLYTRASQPELAPEDEPDLEHHHHHHH</td>
</tr>
<tr>
<td>2</td>
<td>27-50 (24)</td>
<td>signaling peptide</td>
<td>MFQKTYAVFLILLMMFTAACSG</td>
</tr>
<tr>
<td>3</td>
<td>51-380 (330)</td>
<td>modeling unit</td>
<td>SKTSAEKKESETEKSSDIAQVKIKDSYTLPKSYDKSTSDQQLVKVNVA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VKNTGKDPLNVDSDFTLYQGDTKMSDTPEDYSKLQGASTINADKSVETSG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NLFFVVDGKQYELNYTPESYGDKKSIFKIDGKDKKILATADKLQDS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AKALSAYVDFLFGKDNDFEKITGANKNIEVNDFNESAKDGYLASGLS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>STYADSKALDNIVNGIKSLSKSIQAKTTSISKDEAIVEATKVPSAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLSRDIEKVKDYYSKNSASSYEEAVKALQVYPEEFKCLGPASSEKTVE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VKMKNIDQWQLMDDYRAELVEAFIKE</td>
</tr>
</tbody>
</table>
GalaxyTBM server: input page

GalaxyWEB
A web server for protein structure prediction, refinement, and related methods
Computational Biology Lab, Department of Chemistry, Seoul National University

GalaxyTBM
Structure prediction and refinement of up to three loops or termini is automatically performed.

User Information
Job name

E-mail address (Optional)

Sequence Information
SEQUENCE (≤500 AA)

Or SEQ file
(allowed file extensions: fa, fasta, txt, seq)

Submit
submit reset
GalaxyTBM server: output page
GalaxyPepDock server: input page
GalaxyPepDock server: output page
Galaxy7TM server: input page

Galaxy7TM: flexible GPCR–ligand docking by structure refinement

Gyu Rie Lee and Chaok Seok
To whom correspondence should be addressed. Email: chaok@snu.ac.kr

Galaxy7TM
Given a GPCR structure and a ligand structure, optimized complex structures are generated by docking and refinement. Input GPCR structure without gaps in the middle is recommended. Up to five gaps in the input GPCR structure can be filled if its full sequence is submitted together.

User Information
Job name:
E-mail address (Optional):

Input GPCR and ligand structures
PDB File (<1000 AA):
Sequence File (Optional) (<1000 AA):
Ligand File (<150 atoms):

Example
PDB file: P07700.pdb
Ligand file: XF5.mol2
Binding pocket: 99, 90, 93, 178, 179, 246, 250, 290, 273
Report: [View]
Galaxy7TM server: output page
Downloadable GALAXY softwares
(http://galaxy.seoklab.org/softwares)

GALAXY Programs

- GalaxyRefine
- GalaxyPPDock
- GalaxyFill
- GalaxyDock
- FALC
A new energy function (BP2 score) for protein-ligand docking
GalaxyDock with BP2 score
Conclusion: when/how to use computational methods?
Validate/Compare with experimental results

Decide on how much information you would like to use in your computation.

The amount of (experimental) information you use (from literature or your own experiment)

Compromise

Accuracy of your results  New information you obtain
Compare different computational methods

Consensus among the results from different methods

Reliability of your results

Different hypotheses you may test by independent means
Cooperativity and Specificity of Cys$_2$His$_2$ Zinc Finger Protein-DNA Interactions: A Molecular Dynamics Simulation Study

Juyong Lee, Jin-Soo Kim and Chaok Seok*
Department of Chemistry, Seoul National University, Seoul, Republic of Korea

DOI: 10.1021/jp1017289
Publication Date (Web): May 14, 2010
Copyright © 2010 American Chemical Society

* To whom correspondence should be addressed. E-mail: chaok@snu.ac.kr. Phone: 82-2-880-9197. Fax: 82-2-871-8119.

Abstract

My own examples
My own examples

Pulsating Tubules from Noncovalent Macrocycles

Zhegang Huang1,2, Seong-Kyun Kang1, Motonori Banno2, Tomoko Yamaguchi2, Dongseok Seok1, Eiji Yasihara2, Myungsoo Lee1

1Department of Chemistry, Korea University, Seoul 136-701, Korea
2Department of Chemistry, Sungkyunkwan University, Suwon 16419, Korea

To whom correspondence should be addressed. E-mail: myongslee@snu.ac.kr

Science 21 Feb 2017
My own examples
Thank You!