Traditional approaches to 3-D structure determination

X-ray crystallography:
Sufficient protein? Natural protein is often heterogeneous
Can it be expressed? Requires full length expression construct
If so, in soluble state? Often recombinant protein is present as insoluble inclusion bodies
Will it crystallize? Perhaps the most time-consuming task
Will the crystals be of diffraction quality? Some crystals will simply not diffract
How do you solve the “phase” problem? heavy metals, selenomethionine, molecular replacement
NMR spectroscopy:
Same problems above with respect to protein availability.
Protein must be N\textsuperscript{15} and/or C\textsuperscript{13} and/or H\textsuperscript{2} labeled (bacteria must grow in isotopically enriched media, samples may cost > $2,000 to prepare).
Protein must be stable for weeks.
Restriction is terms of size of protein (< 25-30 kD).

All methods are time and labor intensive but yield atomic level resolution (1.5 – 3 Å)
Proliferation of deduced amino acid sequences

1. RTPCR
   
   PCR amplification of RT products using redundant gene specific primers (very rapid; generates large amounts of data)

2. cDNA libraries
   
   Cloning of RT products; when screened yield large numbers of cDNA sequences.

3. Genomic DNA sequencing efforts
   
   Eukaryotes- (man, mouse, chicken, pufferfish, zebrafish, Ciona [tunicate], sea urchin, fruit fly, C. elegans (nematode) and many unicellular pathogenic species)

   Gene finder software extracts EXONs and assembles ORFs.

   Prokaryotes- > 25 genomes have been sequenced
Human genome codes for >30,000 proteins!

Convert sequences to 3-D structure

> 40% sequence identity- HOMOLOGY MODELING (actually extends to much lower degrees of identity)

20-40% sequence identity- THREADING

<25% sequence identity- AB INITIO modeling
Homology modeling- assume query sequence and homologs share the same basic folds; approach is to pick out the best template from a suite of homologs

**Homologous** proteins- share a common evolutionary origin [ancestor]
**Orthologs** - the same gene present in different species (ca., hemoglobin in fish vs. camels)
**Paralogs** - products of gene duplication events (brain, muscle, sarcomeric mitochondrial and ubiquitous mitochondrial creatine kinases in man)

The potential for homology modeling
> 500,000 protein sequences are known
There are >10,000 experimentally determined 3-D structures of proteins
33% of known protein sequences have similarities to a proteins for which the 3-D structure has been determined
Therefore, there is the potential for homology modeling of > 150,000 proteins
Find homologs

1. **Pairwise comparison**: compare target sequence individually with other sequences in a database of sequences; FASTA; BLAST

   Easy to use, on-line sites; probe with protein (BLASTp, tBLASTn) or nucleotides (BLASTn)


   EMB- [http://www.ch.embnet.org/software/BottomBLAST.html](http://www.ch.embnet.org/software/BottomBLAST.html)

2. **Psi-Blast**: constructs multiple sequence alignments of many sequences which are iteratively sampled from a database; position specific scoring matrix is used to sample database for additional homologs; *very useful when percent identity is low*

NCBS, National Center for Biological Sciences, Bangalore

[http://caps.ncbs.res.in/campass/psi-blast.html](http://caps.ncbs.res.in/campass/psi-blast.html)

3. **3D template matching**: pairwise comparison of the query sequence and a protein of known structure; structure-dependent scoring function is used to ascertain whether a query protein adopts or not any one of the library of 3D folds; useful when pairwise or multiple sequence approaches have not identified homologs

Imperial College – 3D PSSM

[http://www.sbg.bio.ic.ac.uk/~3dpssm/](http://www.sbg.bio.ic.ac.uk/~3dpssm/)
4. **PROPSEARCH** uses the amino acid composition instead. In addition, other properties like molecular weight, content of bulky residues, content of small residues, average hydrophobicity, average charge a.s.o. and the content of selected dipeptide-groups are calculated from the sequence as well. 144 such properties are weighted individually and are used as query vector. The weights are then trained on a set of protein families with known structures, using a genetic algorithm. Sequences in the database are transformed into vectors as well, and the euclidian distance between the query and database sequences is calculated. Distances are rank ordered, and sequences with lowest distance are reported on top. (University of Montpelier)

http://www.infobiosud.univ-montp1.fr/SERVEUR/PROPSEARCH/Presentation.html

**Template choice**

1. Higher the sequence identity, the more likely the template will be suitable
2. Most closely related from a phylogenetic point of view
3. Template “environment” (solvent, pH, temperature, quaternary structure)
4. Quality of the template structure (resolution and R factor)
Alignment of target (query) with template- using CLUSTAL, GAP (GCG), BLAST (NCBI) etc with manual adjustments!

Look for highly conserved regions- this helps with the manual adjustments of alignments!

MSTSQNK-IFNQGEDCKFSMTSAGADYNEKUTHLEIMGFVFMERETFYMHLILLAISSK
MTAVAGQFPGVGDCKAFMWDASADTDYETHELEIMGFVFMERETFYMKLAVASCK
--MSAAGP1FSGENHCFQVSHDANADVLEIMGFVFMERETFYMISLAVASCK
--MTSSQ2FIFSQEDCKFETHDANAGYNEKDTEVKEIMGFVFMERETFYMISLAVASCK
--MTAGQIFSQEDCKAGSHDAGYNEKUTHLEILGFVFMERETFYMISLAVASCK
--MSSDK-IFAKAGCKSAHDAGYNETDHELIGFVFMERETFYMISLAVASCK
--SSAAASPLFAEGIDCFPAWSRAAPAYUTSHTLQILGFVFMERETFYMISLAAASER


Match display thresholds for the alignment(s):

| = IDENTITY
|: = 2
|.: = 1

UrechiscaupoLK.txt x limulusak.pep October 29, 2001 12:42
EXAMPLE-
*Chaetopterus variopedatus* (polychaete = marine worm) mitochondrial creatine kinase vs. chicken sarcomeric MiCK (X-ray crystal structure published in 1995 by Fritz-Wolf et al.; PDB-1crk)

**EXAMPLE**

*Chaetopterus variopedatus* (polychaete = marine worm) mitochondrial creatine kinase vs. chicken sarcomeric MiCK (X-ray crystal structure published in 1995 by Fritz-Wolf et al.; PDB-1crk)

**MATCH Display Thresholds for the Alignment(s):**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>_</td>
<td>1 = IDENTITY</td>
</tr>
<tr>
<td>:</td>
<td>= 2</td>
</tr>
<tr>
<td>.</td>
<td>= 1</td>
</tr>
</tbody>
</table>

**correctChaetMiCK.txt** x **chickenMiCKsar.pep** February 25, 2002 14:01

..

```
1 ....................................
MRLGTSNVHLRYPA
14

:  
MAGTFGRLLAGRVTAALFAAAGSGVLTTGYLLNQQNVKATVHEKRKLFPP
50

:  
SANYPDLSQHNNIMASNLTPVIYAKLRDKVTPNGVTLNLCIQTGVDNPGH
64

:  
SADYPDLRKHNNCMAECLTPAIYAKLRDKLTPNGYSLDQCIQTGVDNPGH
80

:  
PFIKTVGLVAGDEESYEVFADLFDKCIDERHGGYKPWD.KHPTDLDSTKL
113

:  
PFIKTVGMVAGDEESYEVFAEIFDPVIKARHNGYDPRTMKHHTDLDASKI
138

:  
RGGNFDPKYVLSSRVRTGRCIRGLSLPPACSRAERREVERVVVEALNGLQ
163

:  
THGQFDERYVLSSRVRTGRSIRGLSLPPACSRAERREVENVVVTALAGLK
190

:  
GDLAGKYFPLAKMTDAEQEKLIEDHFLFDKPVSPLLLASGMARDWPDARG
216

:  
GDLSGKYYSLTNMSERDQQQLIDDHFLFDKPVSPLLTCAGMARDWPDARG
243

:  
IWHNNDKTFLVWINEEDHTRVISMEKGGNMKRVFERFCRGLKEVERLIKE
270

:  
RGWEFMWNERLGYVLTCPSNLGTGLRAGVHVKLPRLSKDPRFPKILENLR
297

:  
LQKRGTGGVDTAAVADVYDISNLDRMGRSEVELVQIVIDGVNYLVDCEKK
324

:  
LQKRGTGGVDTAATGDTFDISNSDRLGKSEVQLVQQVVDGIDLLIQMEKR
351

:  
LERGQ... 376
```

**Generate coordinates for SCRs and VRs using template(s):**

- **Web-based**
  1. Swiss Model- Glaxo Welcome
  2. What If- EMBL

- **Programs**
  1. Modeller- free; Unix, Linux [Rockefeller U.]
  2. Insight II (“homology”)- Unix [MSI]

What If- http://www.cmbi.kun.nl/whatif/


PDB file-header

Background information
PDB file-remarks

**Properties of the structure like resolution, error etc.**

**REMARK 1**

**REMARK** 2
**REMARK** 2 RESOLUTION: 3.0 ANGSTROMS.

**REMARK** 3
**REMARK** 3 REFINEMENT.

**REMARK** 3 PROGRAM: X-PLOR
**REMARK** 3 AUTHORS: BRUNGER

**REMARK** 3
**REMARK** 3 DATA USED IN REFINEMENT.
**REMARK** 3 RESOLUTION RANGE HIGH (ANGSTROMS): 3.
**REMARK** 3 RESOLUTION RANGE LOW (ANGSTROMS): 8.
**REMARK** 3 DATA CUTOFF (SIGMA(F)): 0.
**REMARK** 3 DATA CUTOFF HIGH (ABS(F)): NULL
**REMARK** 3 DATA CUTOFF LOW (ABS(F)): NULL
**REMARK** 3 COMPLETENESS (WORKING-TEST) (%): 99.6
**REMARK** 3 NUMBER OF REFLECTIONS: 40730

**REMARK** 3
**REMARK** 3 FIT TO DATA USED IN REFINEMENT.
**REMARK** 3 CROSS-VALIDATION METHOD: NULL
**REMARK** 3 FREE R VALUE TEST SET SELECTION: NULL
**REMARK** 3 R VALUE (WORKING SET): 0.217
**REMARK** 3 FREE R VALUE: 0.264
**REMARK** 3 FREE R VALUE TEST SET SIZE (%): 10.
**REMARK** 3 FREE R VALUE TEST SET COUNT: NULL
**REMARK** 3 ESTIMATED ERROR OF FREE R VALUE: NULL

---

**PDB file- helix/sheet**

**HELIX** 61 61 ALA D 108 LYS D 110 5 3
**HELIX** 62 62 ARG D 143 GLY D 159 1 17
**HELIX** 63 63 GLY D 162 LEU D 164 5 3
**HELIX** 64 64 GLU D 176 ASP D 184 1 9
**HELIX** 65 65 PRO D 195 ALA D 200 1 6
**HELIX** 66 66 MET D 241 ARG D 262 1 22
**HELIX** 67 67 PRO D 279 ASN D 281 5 3
**HELIX** 68 68 PRO D 295 LYS D 299 1 5
**HELIX** 69 69 PHE D 303 LEU D 310 1 8
**HELIX** 70 70 GLU D 341 GLU D 363 1 23

**SHEET** 1 A 8 GLY A 166 SER A 170 0
**SHEET** 2 A 8 GLY A 211 ASN A 215 -1 N HIS A 214 O LYS A 167
**SHEET** 3 A 8 PHE A 220 ILE A 224 -1 N ILE A 224 O GLY A 211
**SHEET** 4 A 8 THR A 230 LYS A 237 -1 N ILE A 233 O LEU A 221
**SHEET** 5 A 8 VAL A 121 ARG A 130 -1 N ARG A 130 O THR A 230

---

**Secondary structure- [] helices and [ ] sheets**
Three general methods for initial model construction

1. Assembly of rigid bodies- backbone of target is arranged to template positions of carbons

   **RAPPER** (Cambridge University) is a discrete conformational sampling algorithm for restraint-based protein modelling. It has been used for all-atom loop modelling, whole protein modelling under limited (C-alpha) restraints, comparative modelling, ab initio structure prediction, structure validation, and experimental structure determination with X-ray and nuclear magnetic resonance spectroscopy

   [http://raven.bioc.cam.ac.uk/](http://raven.bioc.cam.ac.uk/)
2. Segment matching (coordinate reconstruction)- most
hexapeptide segments of protein structure can be clustered
into approximately 100 structural classes; subset of atomic
positions in template are used as guiding positions to fit
segments from the target into these positions; can be used
to model main chain and side chain atoms

**SegMod** component available in UCLA’s Genemine

[http://www.bioinformatics.ucla.edu/genemine/](http://www.bioinformatics.ucla.edu/genemine/)

3. Satisfaction of spatial restraints-model is built on minimizing
violations of spatial restraints (constraints) as defined from
the template; homology derived restraints (mainchain and
sidechain dihedrals, mainchain CA-CA distances, sidechain-
mainchain distances etc) and stereochemical (bond angles,
dihedral angles, non-bonded atom-atom contacts)

**(MODELLER)**

A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction

A. Fiser, R.K. Do, & A. Sali. Modeling of loops in protein structures,
Initial model will yield a framework of SCRs consisting of the following:

1. Backbone of target protein is arranged to a carbons of the template.
2. Secondary structural elements are present
3. $\phi$ and $\psi$ angles are arranged in SCRs

Initial modeling will yield a framework which lacks two important components -
(a) VRs [non-conserved loops] and (b) side chains.
Loop modeling- changes in loops usually occur in exposed regions that connect secondary structure elements. No reliable methods are available for constructing loops longer than 5 residues. Two general approaches:

1. **Ab initio loop prediction**- conformational search or enumeration of conformations in a given environment guided by a scoring or energy function.

   Also available on the Cambridge University RAVEN site

   [http://raven.bioc.cam.ac.uk/loop2.php](http://raven.bioc.cam.ac.uk/loop2.php)

2. **Database approaches**- Consists of finding a portion of the main chain which fits the stem regions of a particular loop. The database of protein folds is searched and fitted within the constraints of the loop sequence, stem structure and energy optimization criteria.

   Loop database at PKUBIOS (China)

   [http://mdl.ipc.pku.edu.cn/moldes/oldmem/liwz/home/loop.htm](http://mdl.ipc.pku.edu.cn/moldes/oldmem/liwz/home/loop.htm)
Side chains-defined by $\phi_1$, $\phi_2$ .... (dihedral angles)

- protein side chains play a key role in molecular recognition and packing of hydrophobic cores of globular proteins
- significant correlations exist between $\phi_1$, $\phi_2$ and $\psi$ and $\omega$ angles
- side chain conformations exist in a limited number of canonical shapes (rotamers)
- rotamer libraries can be constructed where only 3-50 conformations are taken into account for each side chain
- approaches involve iterative sampling of rotamers for each side chain into the target backbone with scoring matrices and attention paid to steric clashes of sidechain with mainchain.

**Refinement by molecular mechanics**

- restrain the region of the model that is most likely correct and focus on suspect areas or perform on entire molecule
- idealize bond geometry and remove unfavorable non-bonded contacts by protein force fields using
  AMBER Assisted Model Building with Energy Refinement ([http://amber.scripps.edu/](http://amber.scripps.edu/))
  CHARMm Chemistry at HARvard Molecular mechanics ([http://www.accelrys.com/insight/charmm.html](http://www.accelrys.com/insight/charmm.html))
  GROMOS Molecular Dynamics ([http://www.igc.ethz.ch/gromos/](http://www.igc.ethz.ch/gromos/))
-this essentially is an energy minimization approach; during force field calculations there is a tendency for the structure to “drift” away from the control structure. This drift can be reduced by

-(a) limiting the number of minimization steps and

- (b) restraining many of the alpha carbons.

Errors in homology models –(a) errors in side chain packing- as proteins diverge, the packing of side chains in the core changes, (b) distortions or shifts in correctly aligned regions- sequence divergence may result in main chain conformation changes even though overall fold is the same, (c) errors in regions without a template- occurs in segments of a target protein for which there are no equivalents in the template
(d) errors due to misalignments- most common error and (e) incorrect templates- is a problem that occurs when using distantly related templates.

Evaluating models

Explicit approach: overlay the model on the template and evaluate rmsd (root mean square deviation of the $\alpha$ carbons)

- experimental rmsd values by X-ray crystallography range from from 0.5 Å for the same protein to 1 Å for proteins with $> 50\%$ identity

- successful model has $< 2$ Å rmsd from template (if template has a sequence identity $> 60\%$ then the above criterion would be met with a success rate $> 70\%$)

- sequence identity of target and template is critical
Judging quality of homology models

1. **ACCURACY** - how well it fits the templates upon which it was built (rmsd deviation)

2. **CONFIDENCE FACTOR** (Model B-factor) - SWISS MODEL provides a B-factor which is an uncertainty factor; higher the value the lower the amount of actual structural support is present for a particular portion of the model

3. **INCORRECTNESS** - presence of hydrophobic groups on the surface or polar groups in the interior that do not have hydrogen bonding or ionic bonding capabilities satisfied by their neighbors or steric clashes or unreasonable structural parameters such as bond angles/lengths.
4. **Ramachandran diagram**- plot shows main-chain conformational angles; there are a finite number of $\phi$ and $\psi$ angle combinations as defined in the plot; glycines, which lack side chains, often fall out of allowed regions. Residues other than glycines in a model that are not present in allowed regions are suspect.

![Ramachandran Plot](image)

5. **REASONABLENESS**- defined as whether the model is in keeping with expectations for similar proteins. Assessed by summing up the probabilities that each residue should occur in the environment in which it is found in the model. For all PDB models, each of the 20 amino acids has a certain probability of belonging to one of the following classes: solvent-accessible surface, buried polar, exposed nonpolar, helix, sheet or turn. Regions in the model that do not fit expectations are suspect.

Threading Energy- a criterion for reasonableness; higher the energy, lower the reasonableness
Detection of errors
1. manual inspection
2. checking stereochemistry
   - PROCHECK (www.biochem.ucl.ac.uk)
   - WHATCHECK (www.sander.embl-heidelberg.de)
   - SQUID (www.yorvik.york.ac.uk)
3. statistical analyses of structures based on compiled 3-D features of many proteins
   - VERIFY3D (www.doe-mbi.ucla.edu)
   - ProsaII (www.came.sgb.ac.at)
SwissModel First Approach Mode

Please fill these fields:

Your Email address: [Field] (MUST be correct!)
Your Name: [Field]
Request title: [Field] Will be added to the results header

REMEMBER: I want to save my name and email for next login

Provide a sequence or a SWISS-PROT AC code:

[Input field]

NOTES: A SWISS-PROT AC code looks like this: P04406
Sequences can be provided in either RAW, SWISS-PROT, FASTA or GCG format.

Now: [Send Request] or [Reset Form]

Options:

- Define the lower BLAST P(N) limit for template selection:
  
  Lower BLAST limit: [Field] 0.00001

- Define the templates you wish to use for this request:

  In some cases you may wish to define a set of template structures to be used for a modelling attempt. As an example, modelling a serpin (Serine esterase inhibitor similar to the plasminogen activator inhibitor, antithrombin III, etc) will generally fail, since these proteins have two distinct structures in the database:
  1. the activated form of all true serpins
  2. the precursor form as found in the serpin-analogue Ovalbumin.

  It is thus best to choose the correct template(s) you wish to base your model on.
  To do so you may use ExPDB templates, your own templates, or a combination of both.

  - Using ExPDB templates:
    1. Search for suitable templates in the ExNRL-3D database.
    2. Select the entries you consider appropriate from the hit list and/or check if their codes exist in the ExPDB database.
    3. List up to 5 entries in the window below, separated by space.

    NOTE: The ExPDB database is derived from the PDB database. Each chain is in a separate file and the residues have been renumbered continuously. The ExPDB codes are built from the PDB codes as explained above.
Results options: (The SWISS-MODEL server will return all results via Email)

- **Swiss-PdbViewer mode**: will return the final model and the template(s) as a Swiss-PdbViewer project file and a log file tracing all actions taken by the server.
- **Normal mode**: includes the final model coordinates file in PDB format and a log file tracing all actions taken by the server.
- **Short mode**: will return only the final model coordinates file.

- Send the results **as plain ASCII mail** instead of email attachment.
- Include a [WhatCheck report](#) of the final model.

---

**SwissModel Optimise Mode**

Please fill these fields:

- **Your Email address**: [Enter] (MUST be correct!)
- **Your Name**: [Enter]
- **Request title**: [Enter] Will be added to the results header.

Provide a SWISS-MODEL project file:

[Now] Send Request [Reset Form]
correctChaetMiCK.txt x chickenMiCKsar.pep February 25, 2002 14:01 ...

1 ....................................
2 MRLGTSNVHLRYPA
3 ....................................
4 MAGTFGRLLAGRVTAALFAAAGSGVLTTGYLLNQQNVKATVHEKRKLFPP
5 ....................................
6 SANYPDLSQHNNIMASNLTPVIYAKLRDKVTPNGVTLNLCIQTGVDNPGH
7 ....................................
8 SADYPDLRKHNNCMAECLTPAIYAKLRDKLTPNGYSLDQCIQTGVDNPGH
9 ....................................
10 PFIKTVGLVAGDEESYEVFADLFDKCIDERHGGYKPWD.KHPTDLDSTKL
11 ....................................
12 PFIKTVGMVAGDEESYEVFAEIFDPVIKARHNGYDPRTMKHHTDLDASKI
13 ....................................
14 RGGNFDPKYVLSSRVRTGRCIRGLSLPPACSRAERREVERVVVEALNGLQ
15 ....................................
16 THGQFDERYVLSSRVRTGRSIRGLSLPPACSRAERREVENVVVTALAGLK
17 ....................................
18 GDLAGKYFPLAKMTDAEQEKLIEDHFLFDKPVSPLLASGMARDWPDARG
19 ....................................
20 GDLSGKYYSLTNMSERDQQQLIDDHFLFDKPVSPLLTCAGMARDWPDARG
21 ....................................
22 IWHNDKKNFLVWVNEEDHTRVISMQKGGNMREVFDRFCNGLQKVENLIQS
23 ....................................
24 IWHNNDKTFLVWINEEDHTRVISMEKGGNMKRVFERFCRGLKEVERLIKE
25 ....................................
26 RGWEFMWNEHLGYVLTCPSNLGTGLRAGVHIKIPKLAKDPRLNEVLKKMN
27 ....................................
28 RGWEFMWNERLGYVLTCPSNLGTGLRAGVHVKLPRLSKDPRFPKILENLR
29 ....................................
30 LQKRGTGGVDTAATGDTFDISNSDRLGKSEVQLVQQVVDGIDLLIQMEKR
31 ....................................
32 LQKRGTGGVDTAAVADVYDISNLDRMGRSEVELVQIVIDGVNYLVDCEKK
33 ....................................
34 LERGQRIDDLIPK
35 ....................................
36 LEKGQDIKVPPPLPQFGRK
37 ....................................
38

Chicken sarMiCK
Where do we go from here?

- High through-put structure determination efforts will never come close to determining the bulk of structures.
- However, a dictionary of most of the possible folds is a realistic.
- Given these folds, homology modeling affords great promise in identification and characterization of structural genes whose products have an unknown function.