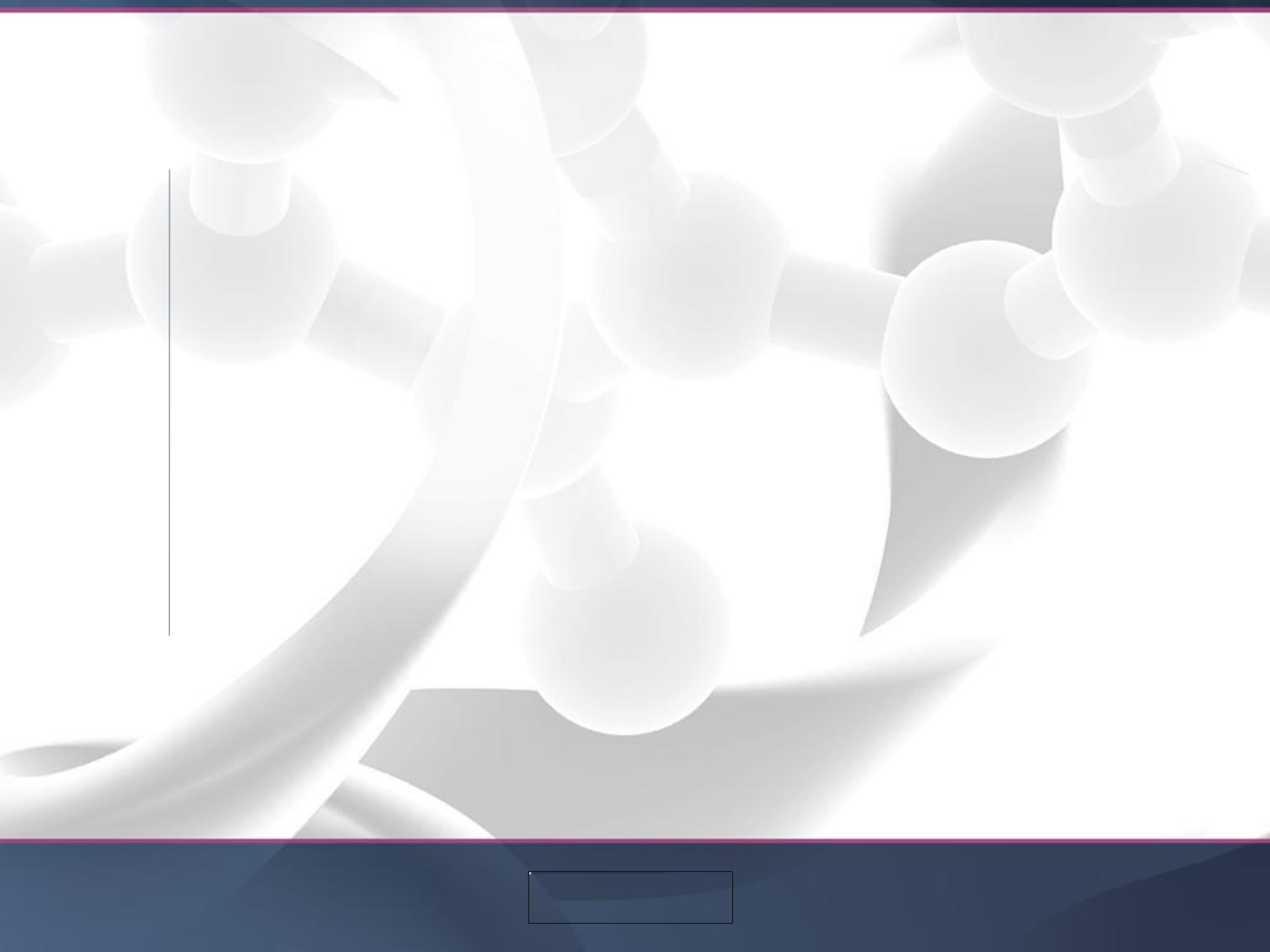
Homology Modeling with Discovery Studio

Protein Homology Modeling

Anand Krishnamurthy, Ph.D.
Senior Scientist- Life Sciences, Modeling & Simulations
akrishnamurthy@accelrys.com

June, 2013



But is sequence enough?

With sequence searching alone:

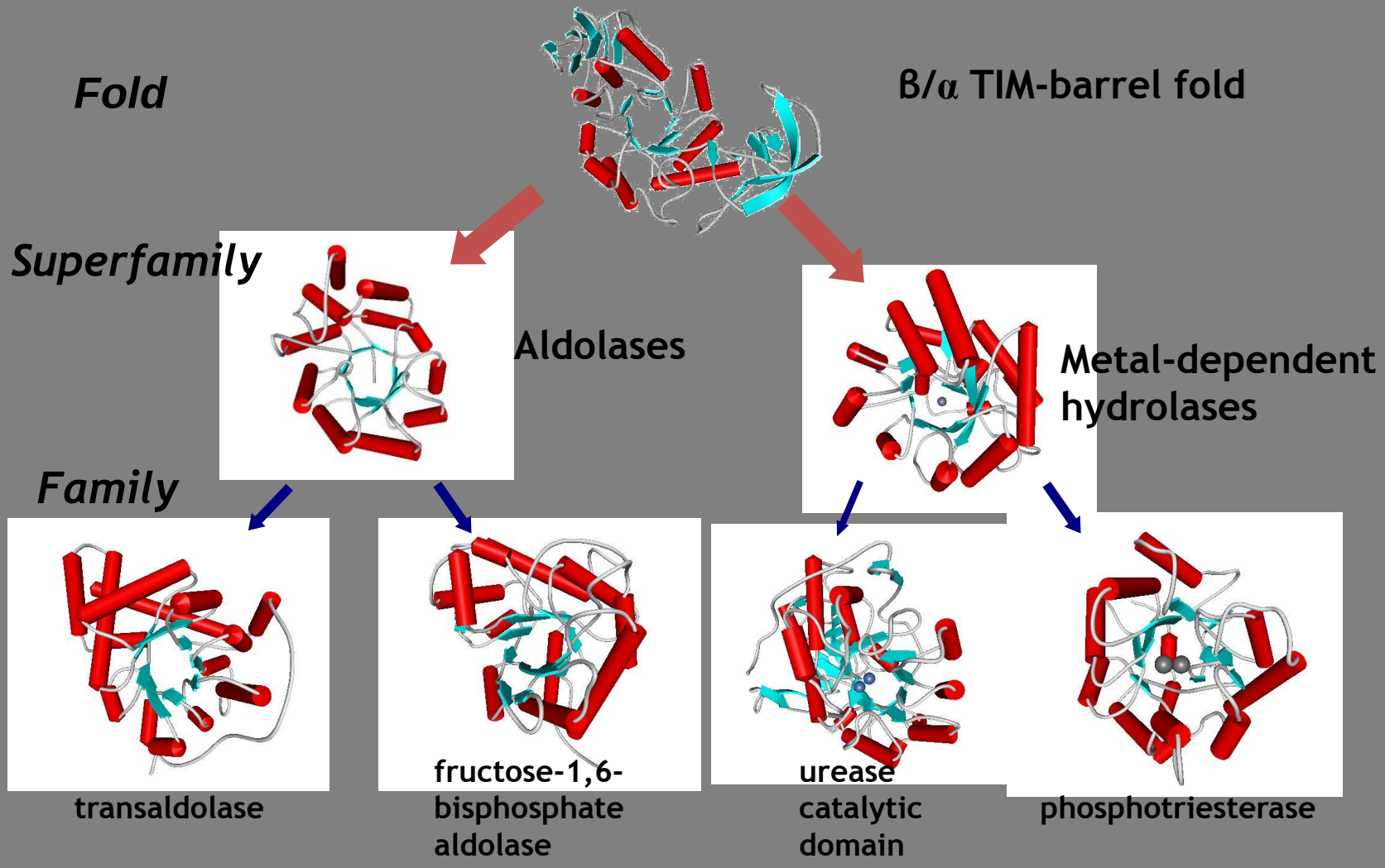
- Is it possible to identify all members of a divergent family of proteins based solely on the proximity of their sequences?
- Can sequence alone identify folds of proteins?
- Can sequence alone divulge information about the organic chemistry occurring at the active site?

The key to understanding protein function is structure.

Where to Obtain Structures

- Protein Data Bank
  - URL http://www.rcsb.org/pdb
  - Run by the Research Collaboratory for Structural Bioinformatics (RCSB)
- Fold Classification Databases
  - SCOP
    - http://scop.mrc-lmb.cam.ac.uk/scop/
    - Created largely by manual inspection
  - CATH
    - http://www.cathdb.info
    - Created through automatic methods with manual inspection and verification

SCOP Levels of Classification

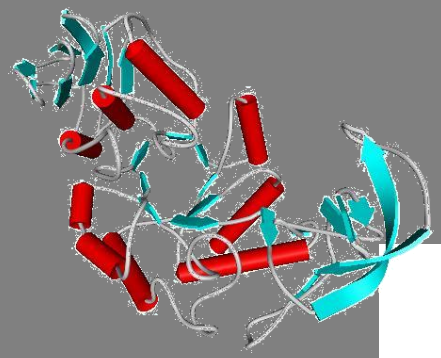


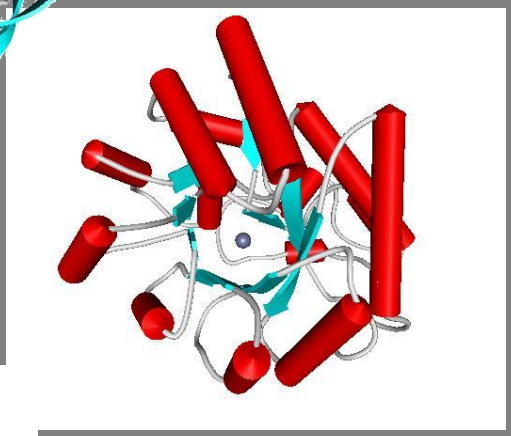
How many structures to consider?

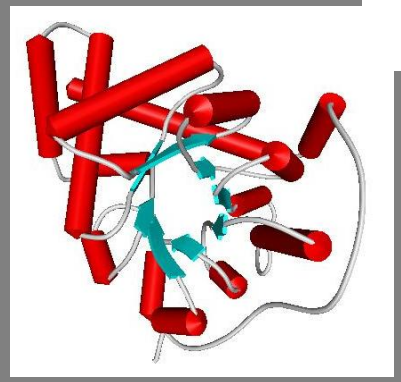
- The Protein Data Bank contains over 82,160 structures
  - As of June 2012
  - Mostly redundant
    - At 95% sequence identity, there are over 18,000 clusters
- Over 1,195 protein folds have been identified so far
  - As of the SCOP release 1.75, February 2009
- Total number of protein folds is thought to be finite
  - Coulson and Moult (2002) estimate 10,000 folds

So use the structural redundancy!

- Similar sequences fold into similar structures.
- Also similar structures may occur due to...
  - Similar evolutionary origins
  - Similar functions
  - Similar folding patterns
- We can exploit the structural relationship!



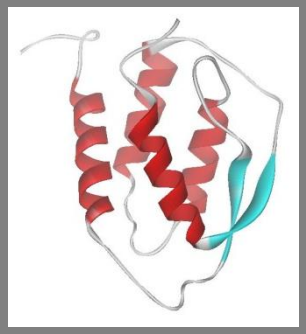


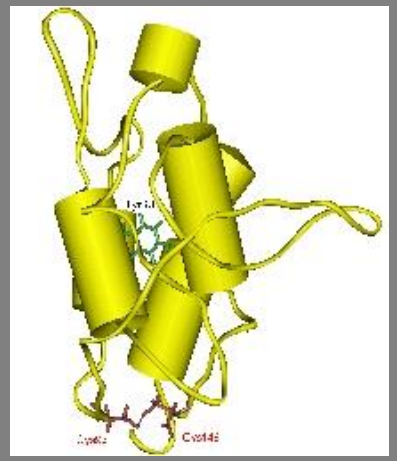


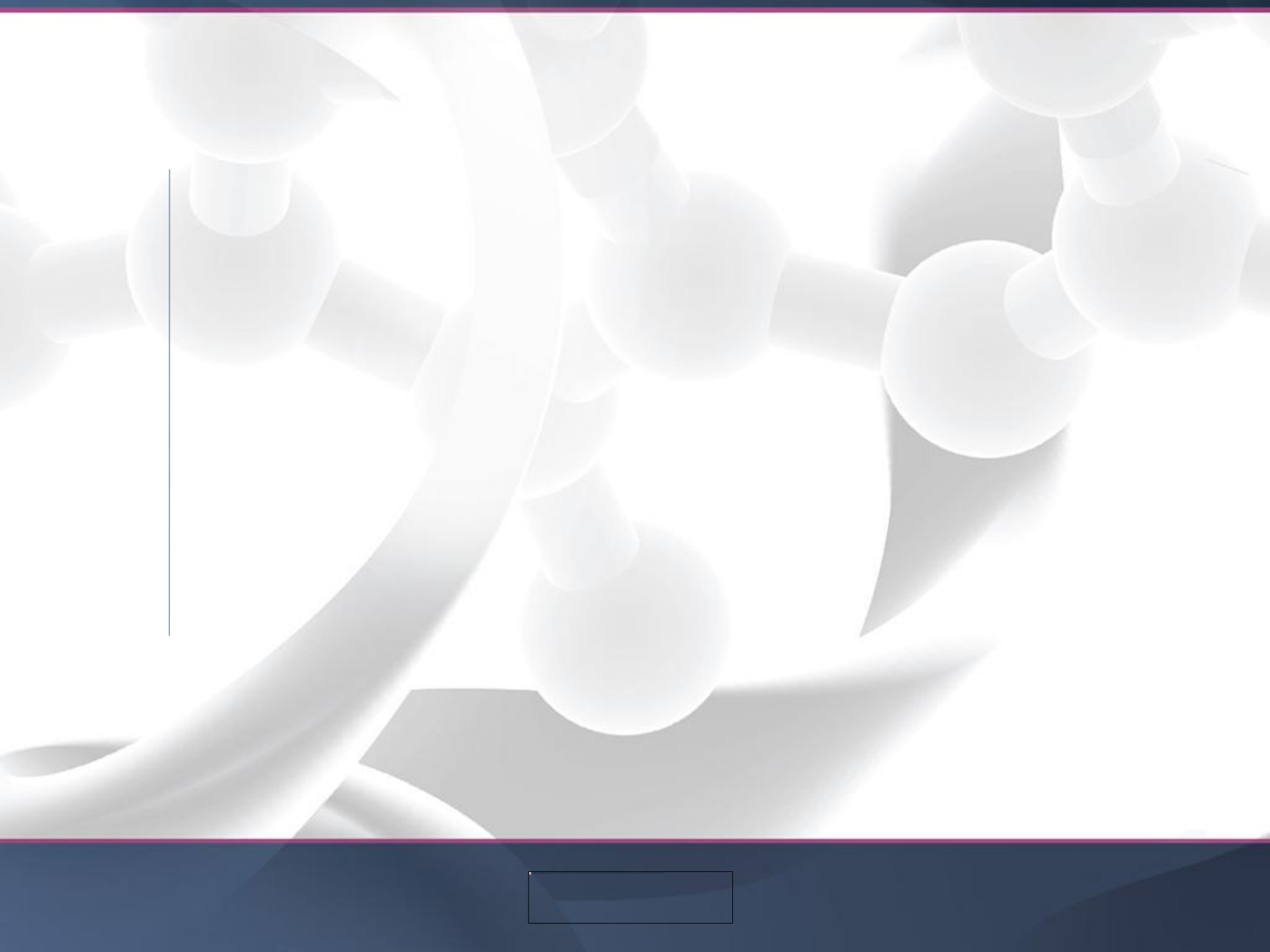
Homology Modeling

- Take the sequence of an unknown structure
- Map it onto a known structure
- Yield a theoretical model

GTLCGFLW…

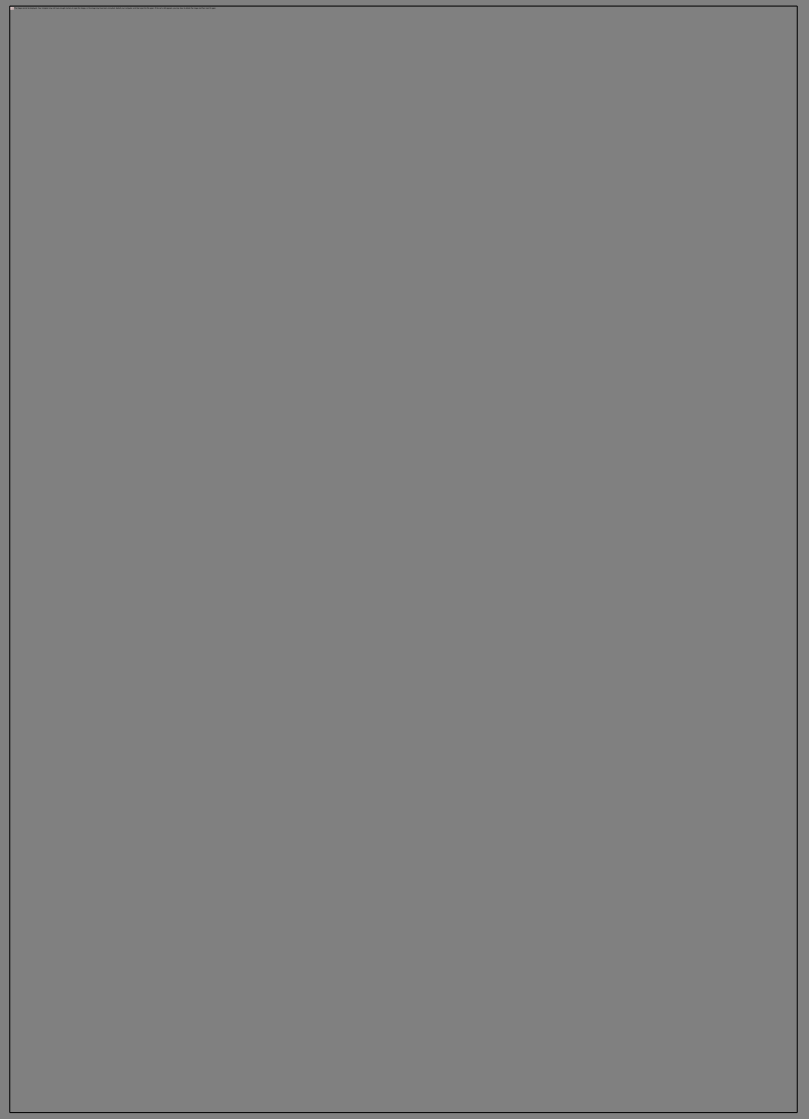






Homology Modeling Examples

- Lee and Briggs (2004)
- Aminoacyl-tRNA synthetases (aaRSs)
  - Catalyze binding of a specific amino acid to tRNA
  - Accomplished via proofreading and editing mechanisms
- Mursinna et al. (2001) reported in *E. coli* leucyl-tRNA synthetase (LeuRS)
  - Substitution of a highly conserved threonine (T252) with an alanine within the editing domain
  - Caused LeuRS to cleave its cognate aminoacylated leucine from tRNALeu



Homology Modeling Examples

- Lee and Briggs (2004)
- Homology model of *E. coli* LeuRS was constructed
  - Used an X-ray crystal structure of *Thermus thermophilus* LeuRS
- Docked leucine ligand into the editing domain
- Identified two possible amino acid binding sites
  - One near the highly conserved T252 and is important for amino acid discrimination
- Mursinna et al. (2004) confirmed T252 side chain
  - Sterically blocks charged leucine from editing site
  - Also, maintains proper geometry of protein

Quality of Model Related to Quality of Template

- Assignment of function/protein fold
- Finding binding or active sites by 3D motif searching
- Success may require additional information

- Identify properties that do not occur in template (e.g. electrostatics, cavity volume)
- Could be used for protein engineering (e.g. stability)
- May help guide site-directed mutagenesis studies

- Comparable to medium resolution X-ray/NMR structure
- May be useful for structure-based design
- Applicable for docking studies
- May help understand substrate or ligand specificity

Requirements for a Homology Project

- Absolute requirements
  - The amino acid sequence of the protein to be built
    - The sequence of the unknown or target protein
  - The high-resolution structure of a related protein
    - The reference structure or template
- Optional information
  - Any additional reference protein structures
    - Multiple related structures
    - To reinforce the alignment of the unknown based upon structural consideration
  - Additional sequences of related proteins
    - To strengthen the alignment of the unknown based on sequence alignment

Homology Modeling Assumptions

- The overall 3-D structure of the unknown protein is similar to that of the related proteins.
  - Greater confidence at higher levels of sequence identity
  - Otherwise, need strong experimental data to support assumption
- Regions of homologous sequence have similar structure.
  - Similar sequences yield similar structures.
- Residues homologous throughout a family of proteins are conserved structurally.
  - Could be conserved for a variety of reasons, e.g. active site residues.

Homology Modeling Assumptions

- Residues involved in biological activity have similar topology throughout the protein family.
  - The topology dictates the organic chemistry
- Loop regions (non-conserved residues) allow insertions and deletions without disrupting the overall structure of the protein.
  - Loops are more tolerant of changes while retaining the overall fold of the protein.
- Loop regions are flexible and therefore need not be constructed as strictly as the conserved regions…

  unless they play a role in biological activity.

Critical Evaluations of Homology Modeling

- CASP - Critical Assessment of techniques for protein Structure Prediction
  - Biennial meeting with the first meeting in 1994
  - An in-depth, objective, and blind assessment of current abilities and inabilities in the area of protein structure prediction
  - Results and information posted at http://predictioncenter.org/

- Aims of CASP
  - Determine the state of the art methods in structure prediction
  - Identify progress
  - Reveal bottlenecks
  - Show where effort may best be focused

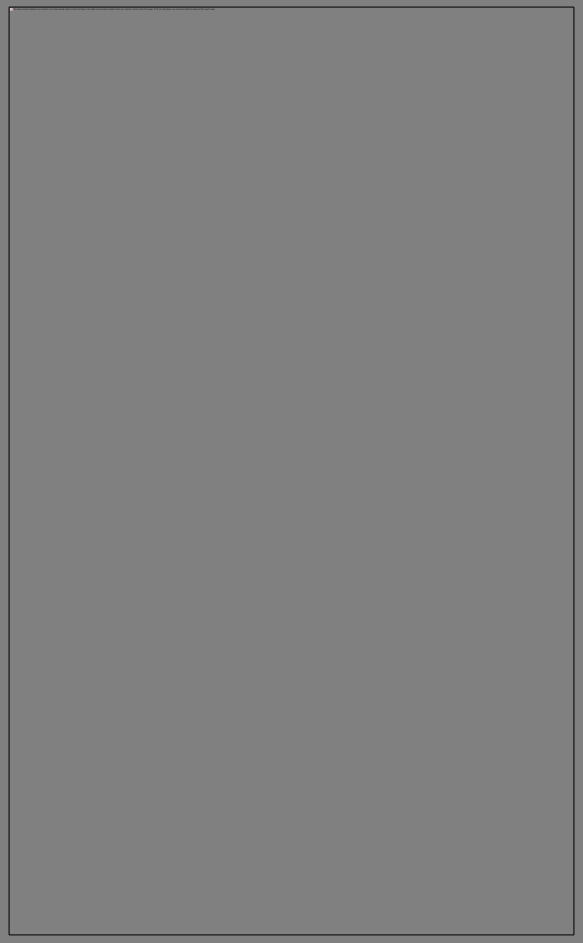
Critical Evaluations of Homology Modeling

- Organizers of CASP issue a call for targets
  - X-ray crystallography groups are asked to defer publication of structures
- Information is provided to CASP participants about target
  - Protein name
  - Organism
  - Amino acid sequence
  - Additional pertinent information such as domains or multimeric complex
- Participants attempt to predict structure
  - Use any technique available
- Results are assessed by CASP organizers
  - Presented at the meeting with the experimental structures

Critical Evaluations of Homology Modeling

- CASP8 predicted model classification categories
  - Template based modeling
    - Comparative modeling
    - Homology modeling
    - Some “easy” analogous fold-based modeling
  - Template free modeling
    - Models not derived from templates
    - Proteins with novel folds
    - Difficult analogous fold-based models

- Results provide a wealth of data concerning techniques and methodologies



And now for today’s problem...

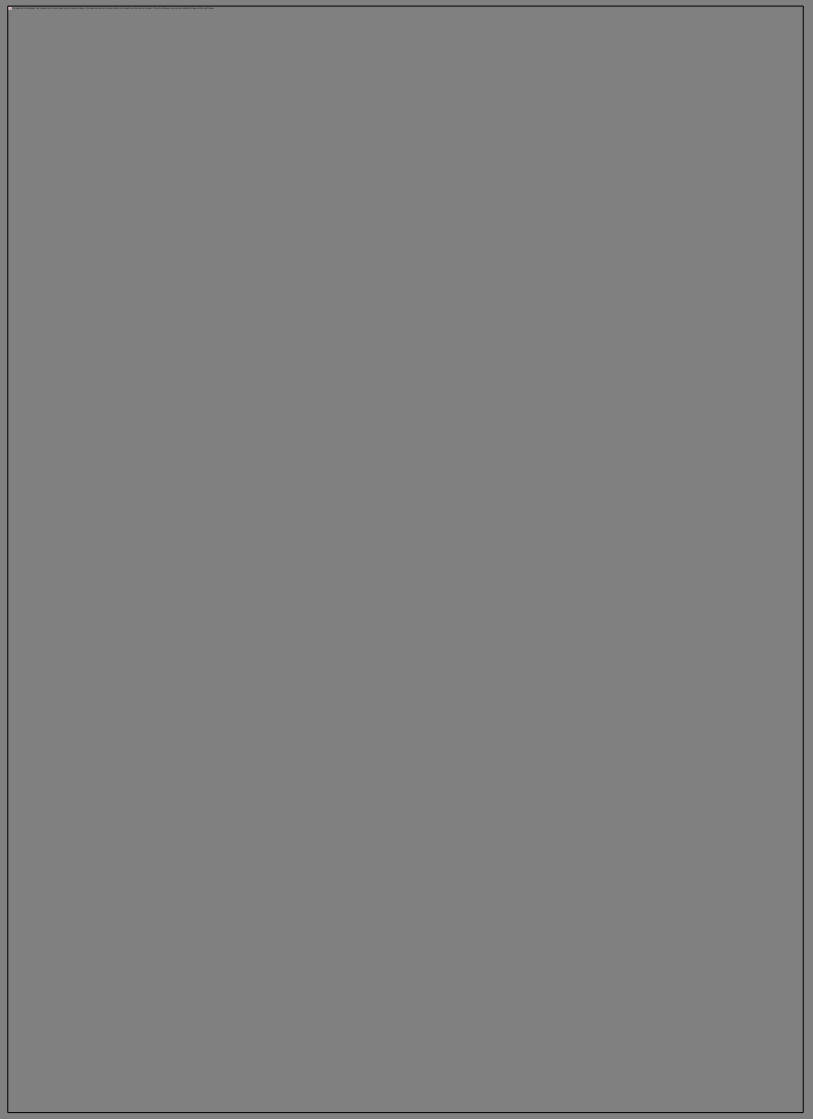
Histone Deacetylases

- Acetylation in histones
  - Play a crucial role in transcription and gene regulation
  - Occurs at the ε-amino group of specific lysines
  - Occurs via two groups of proteins
    - Acetyl transferases
    - Histone deacetylases (HDACs)
- Inhibition of histone deacetylases induce in cancerous cells
  - Growth arrest
  - Differentiation
  - Apoptotic cell death

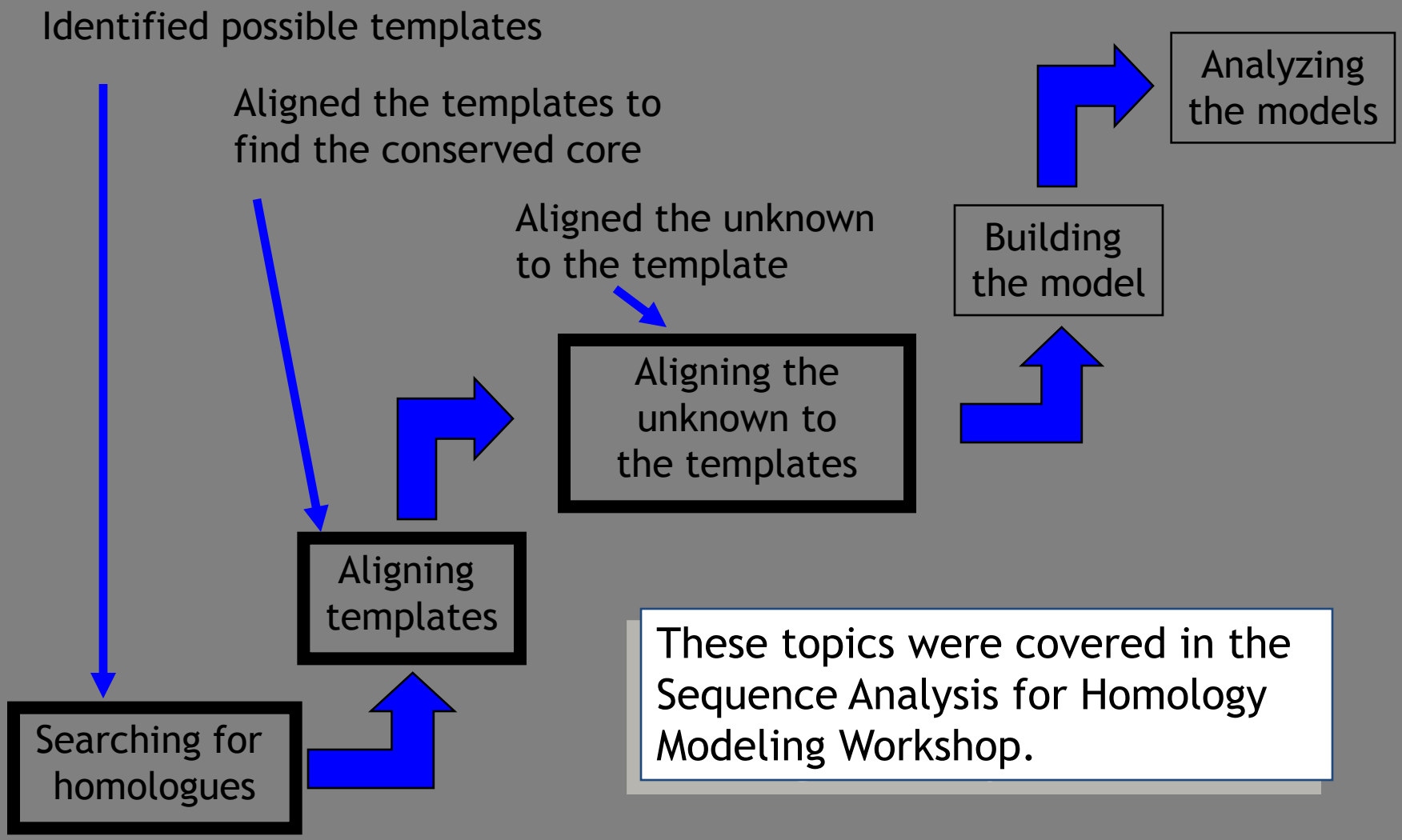
Histone Deacetylases

- Possible target for anticancer drugs
  - Several classes of HDACs with distinct roles in gene expression
  - Specific inhibitors are desired
- Several classes of inhibitors are known

- Wang and co-workers (2005)
  - Used homology modeling to build models of four class 1 HDACs
  - Used models to identify features of binding sites

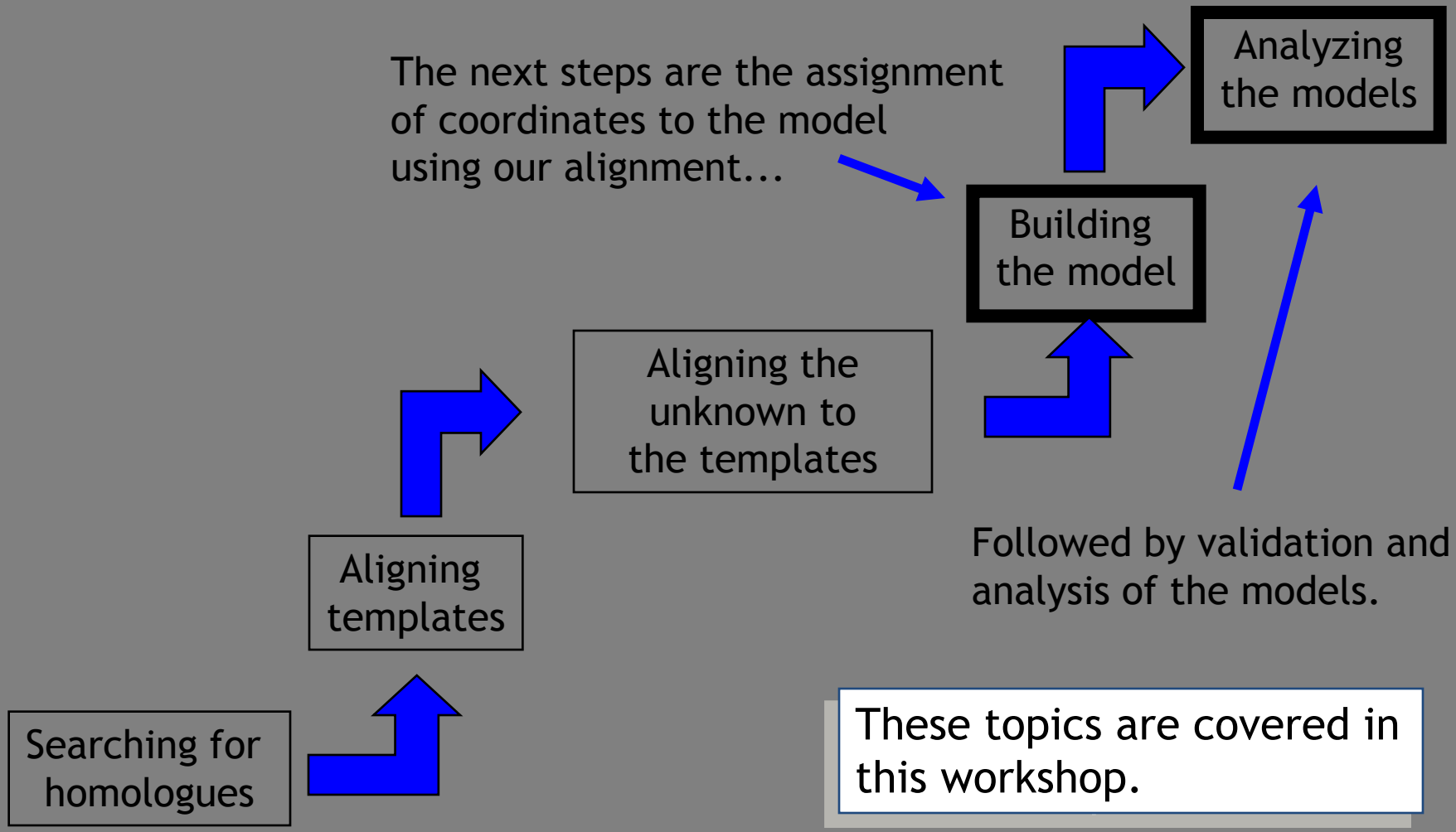


Structure Prediction by Homology Modeling



These topics were covered in the Sequence Analysis for Homology Modeling Workshop.

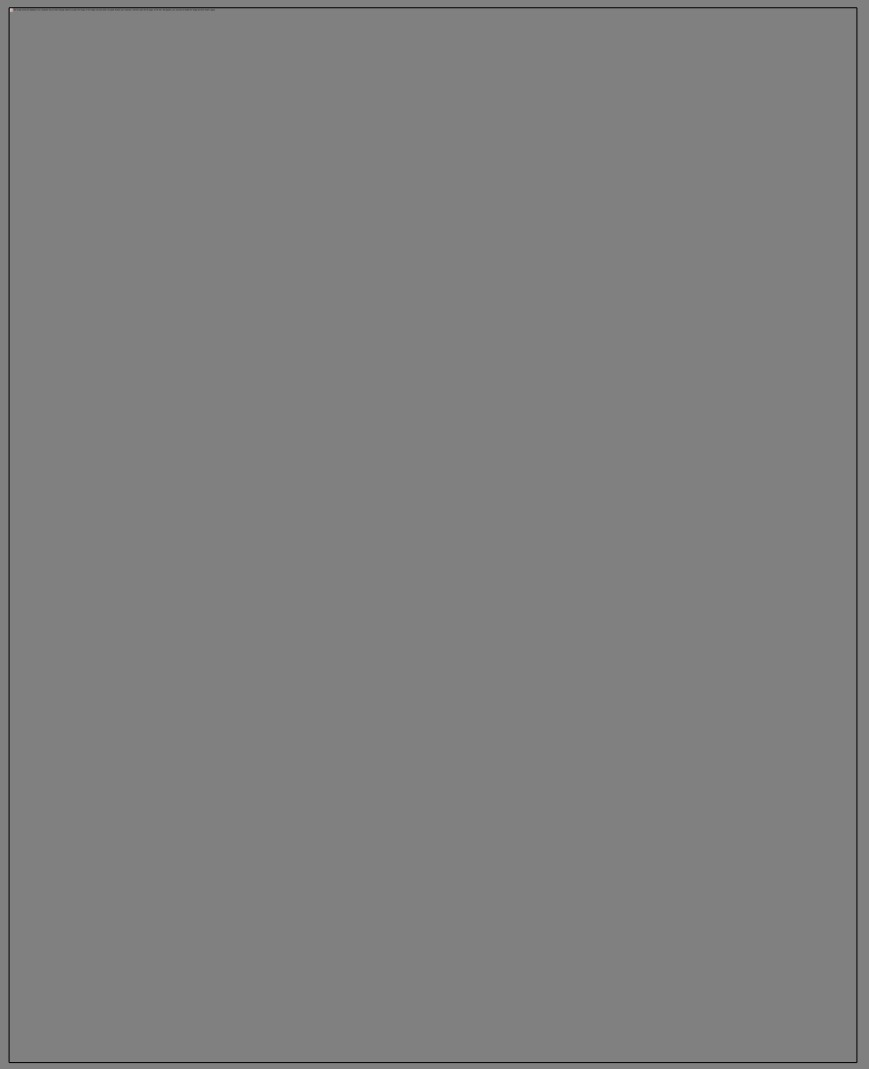
Structure Prediction by Homology Modeling



In this workshop…

- Use Discovery Studio to build a model of histone deacetylase HDAC1

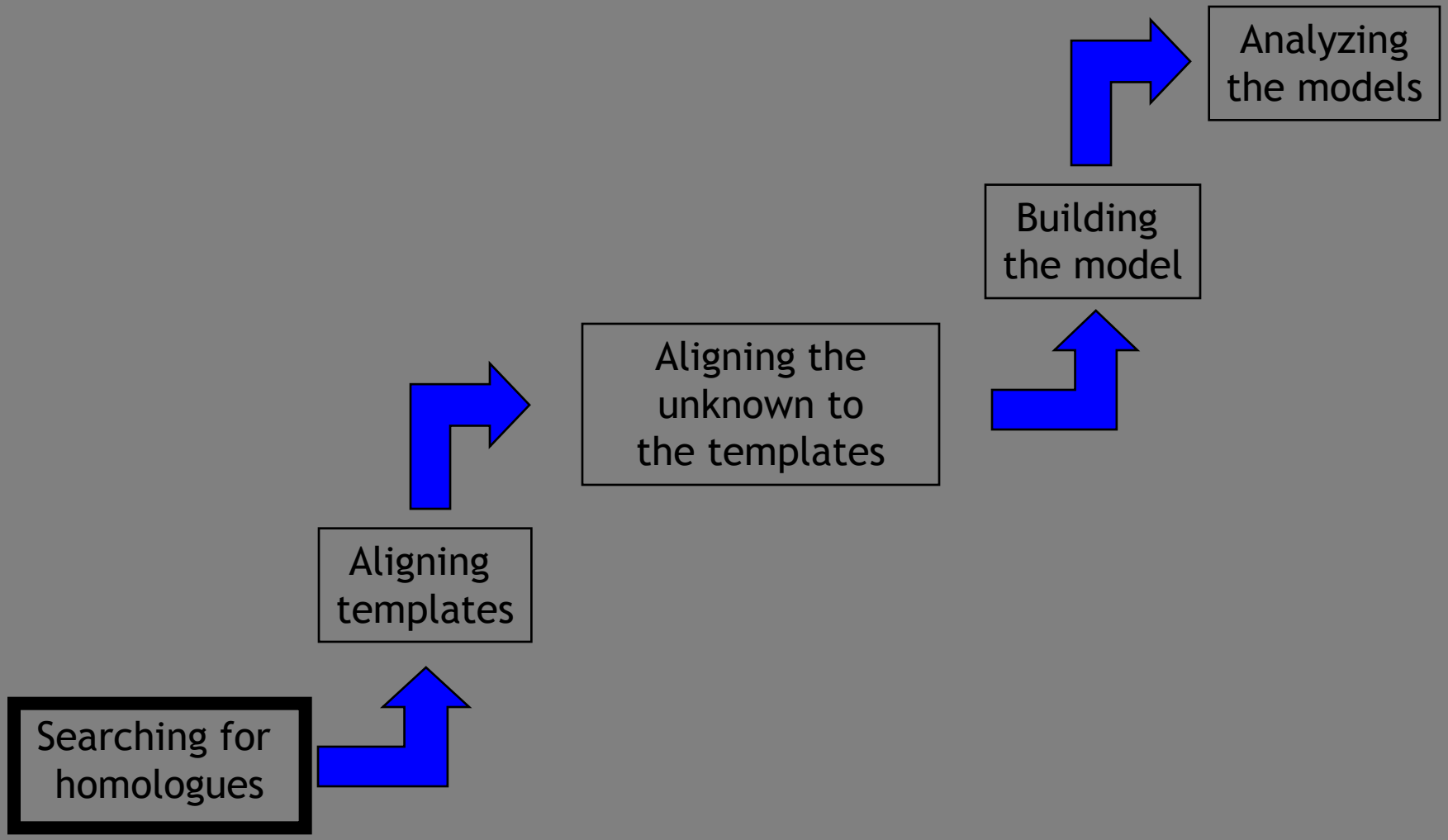
- Use multiple templates
  - Two known homologues
  - Use a prepared alignment
- Generate three models with MODELER
- Evaluate the models
- Refine a model



Review of Template Identification and Alignments

Finding Sequence and Structural Homologues

Structure Prediction by Homology Modeling



Importing Query Sequences

- The query sequence must be brought into Discovery Studio
  - Regardless of where the query sequence is obtained

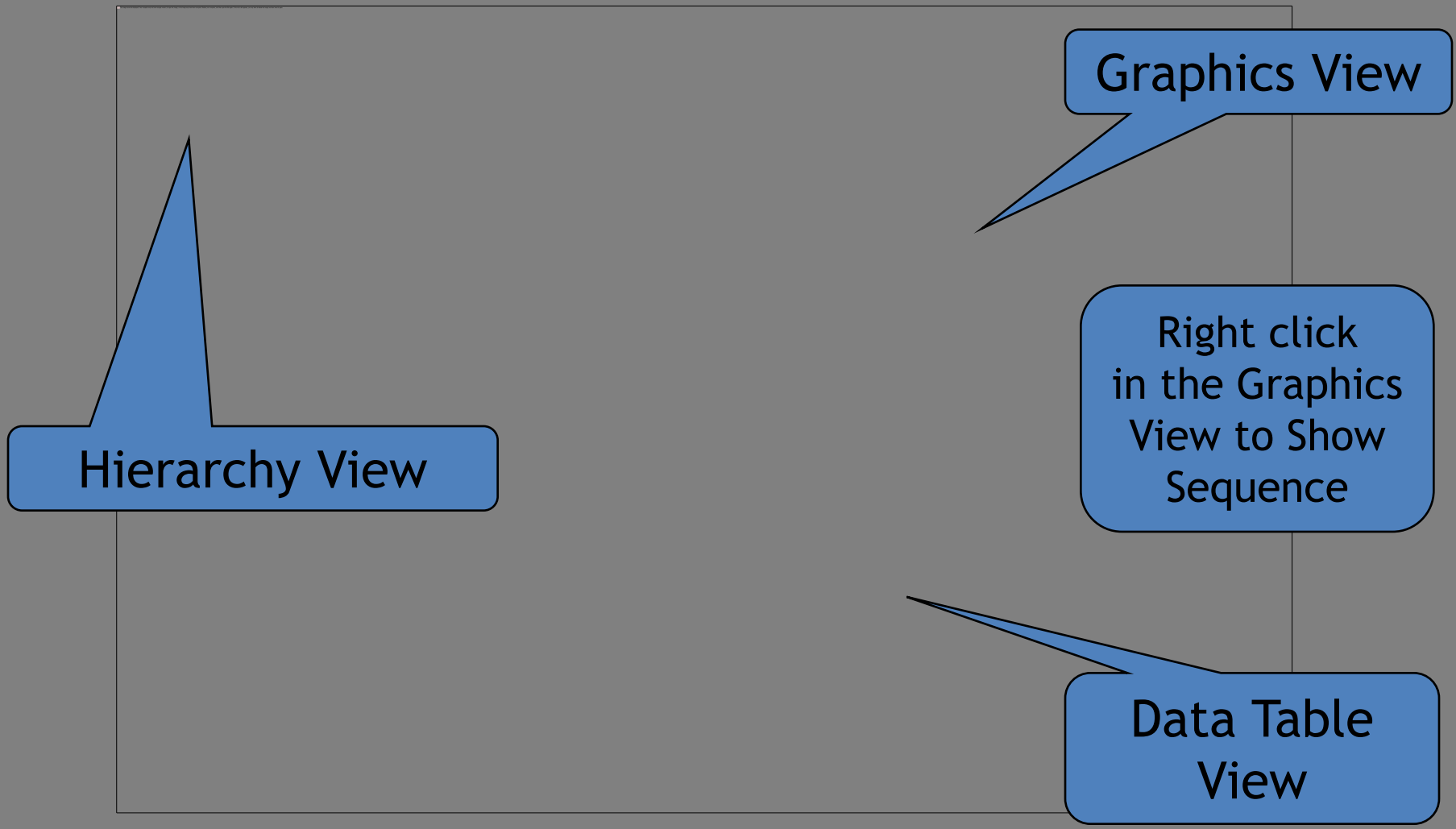
- Supported sequence file formats:

  - Bioinformatic Sequence Markup Language (.bsml)
    - Default
  - Biosym format (.seq)
  - Pearson (FASTA) format (.fasta)
  - NBRF-PIR format (.pir)
  - Swiss-Prot format (.sws)
  - GenBank (.gb)
  - ClustalW (.aln)
  - Biosym Alignment (*.align)
  - Clustal (*.aln)
  - GDE (*.gde)
  - GCG format (.ssf and .rsf)
  - GCG multiple sequence (*.msf)

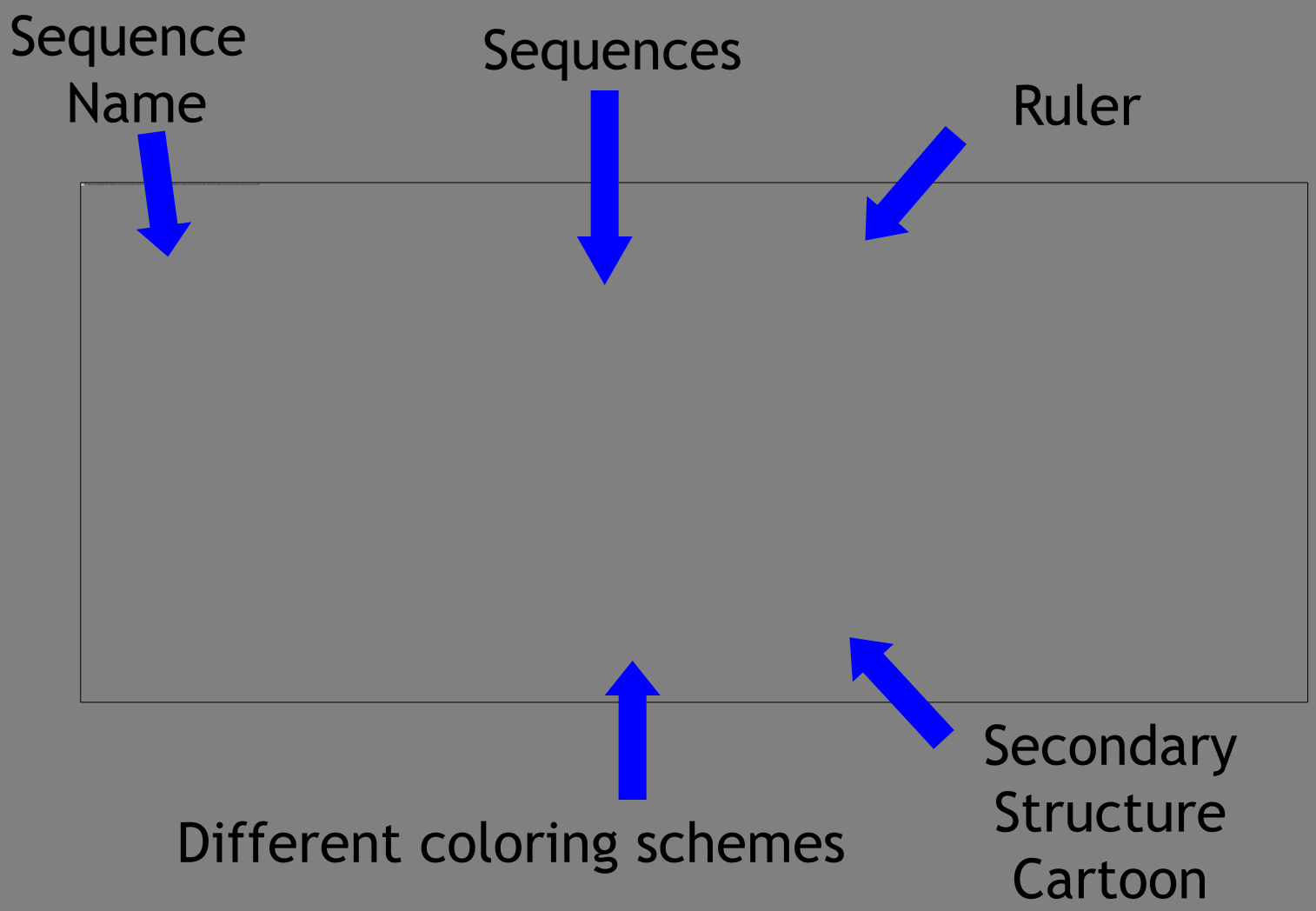
Importing Query Sequences

- Single and multiple sequence file formats may be read from a file
  - **File | Open…**
  - **File | Insert From…**
- Copy and paste into an empty Sequence Window
  - **File | New | Protein Sequence Window**
- Single sequences may also be loaded through the **File | Open URL...** command

Sequence Annotation Window

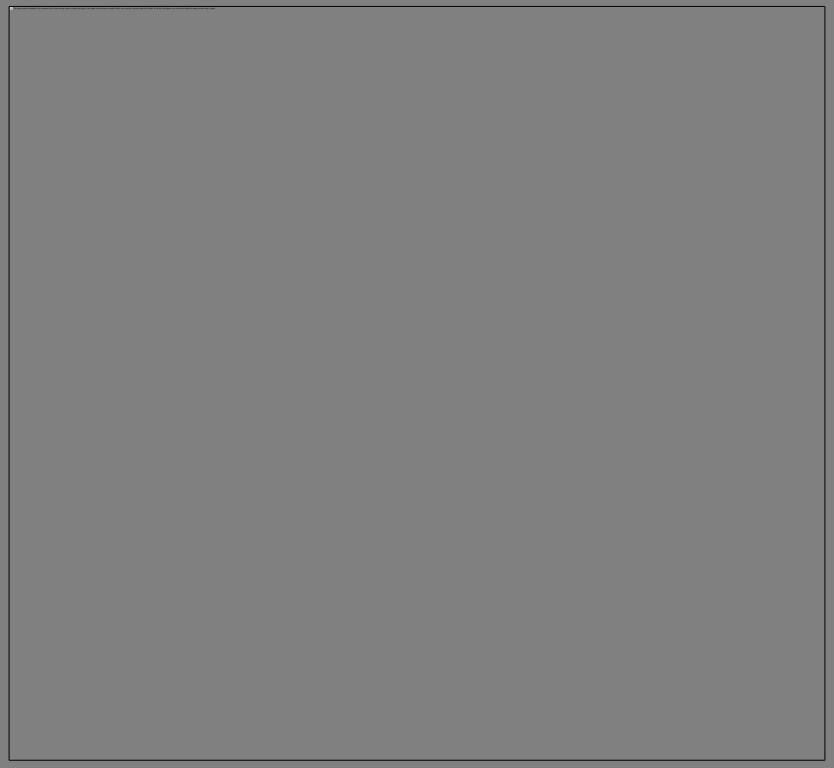


The Sequence Window



Search Protocols

- All found in the Sequence Analysis group
  - BLAST
    - Runs the BLAST protocol on local database
  - NCBI-BLAST
    - Runs the BLAST on the NCBI server
    - Searches against the database provided on the NCBI server
  - PSI-BLAST
    - Runs the PSI-BLAST protocol on a local database
  - Scan Sequence Profiles
    - Compares an input sequence profile against a sequence profile database



Comparison of Database Searches

- Structural Databases
  - Example: PDB_nr95
  - Identify three-dimensional homologues to unknown
  - A hit is required for building the homology model
  - Multiple hits can reinforce hypotheses
    - Function
    - Structural family
    - Important residues
  - Valid hits cover a folding domain
    - May not be your entire unknown sequence
- Sequence Databases
  - Examples: SWISS-PROT or NCBI
  - Identify sequence homologues
  - Depending on database there may not be an available structure
  - Multiple hits can
    - Create a PSSM that can be used to search the structural databases
    - Reinforce a sequence alignment of unknown and templates
    - Aid in function prediction

Databases in Discovery Studio

- Search locally or at the NCBI database
  - At NCBI, only ungapped BLAST or BLAST
- PDB
  - Derived from PDB database maintained at RCSB
  - Each entry is a single chain
  - Hits are associated with structures
- PDB_nr95
  - Only high resolution and NMR structures
  - Any two sequences in the database have less than 95% sequence identity
  - Hits are associated with structures

Databases in Discovery Studio

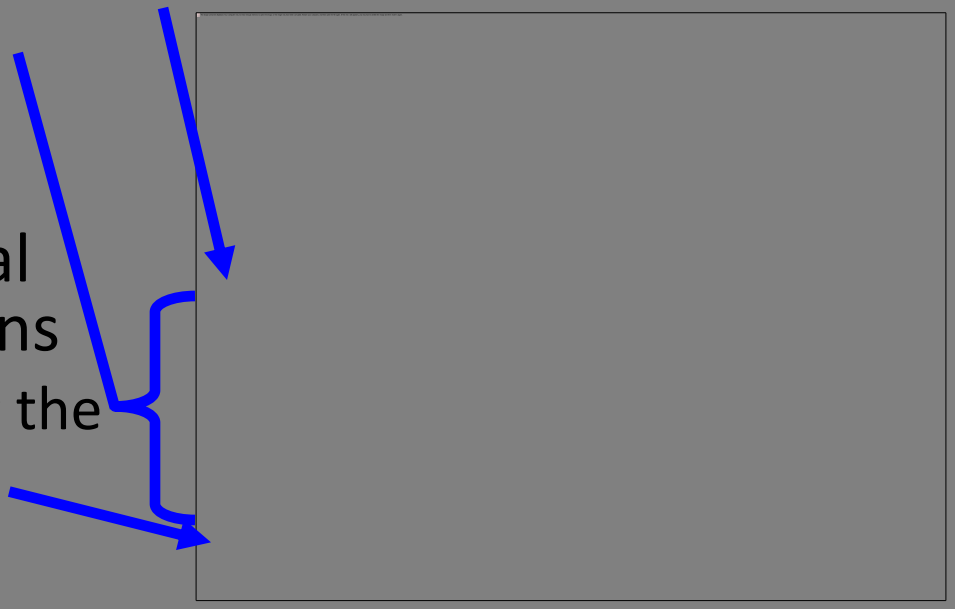
- SwissProt
  - Curated protein sequence database with annotations
  - Maintained by Swiss Institute of Bioinformatics
  - Does not include sequences in TrEMBL (a supplemental database to SwissProt)
- UniRef90
  - A non-redundant protein sequence database from UniProt
  - Any pairwise sequence identity is less than 90%
  - Recommended for PSI-BLAST sequence searches to develop a robust position specific substitution matrix
- NRDB
  - From NCBI
  - Contains non-redundant protein sequences from GenBank, CDS translations, PDB, SwissProt, PIR, and PRF databases.

Running a local BLAST Search

- Use the BLAST Search (DS Server) protocol
  - Found in the Sequence Analysis folder
- Specify the sequence window and the query sequence
- Specify the database to search
  - Databases located on the server
- Choose the scoring matrix

Running a local BLAST Search

- Use the BLAST Search (DS Server) protocol
  - Found in the Sequence Analysis folder
- Gap penalities are optimized for the specific scoring matrix selected
- Can specify options to fine tune the search
- Option to include additional BLAST command line options
  - See the Parameter Help for the list of available option

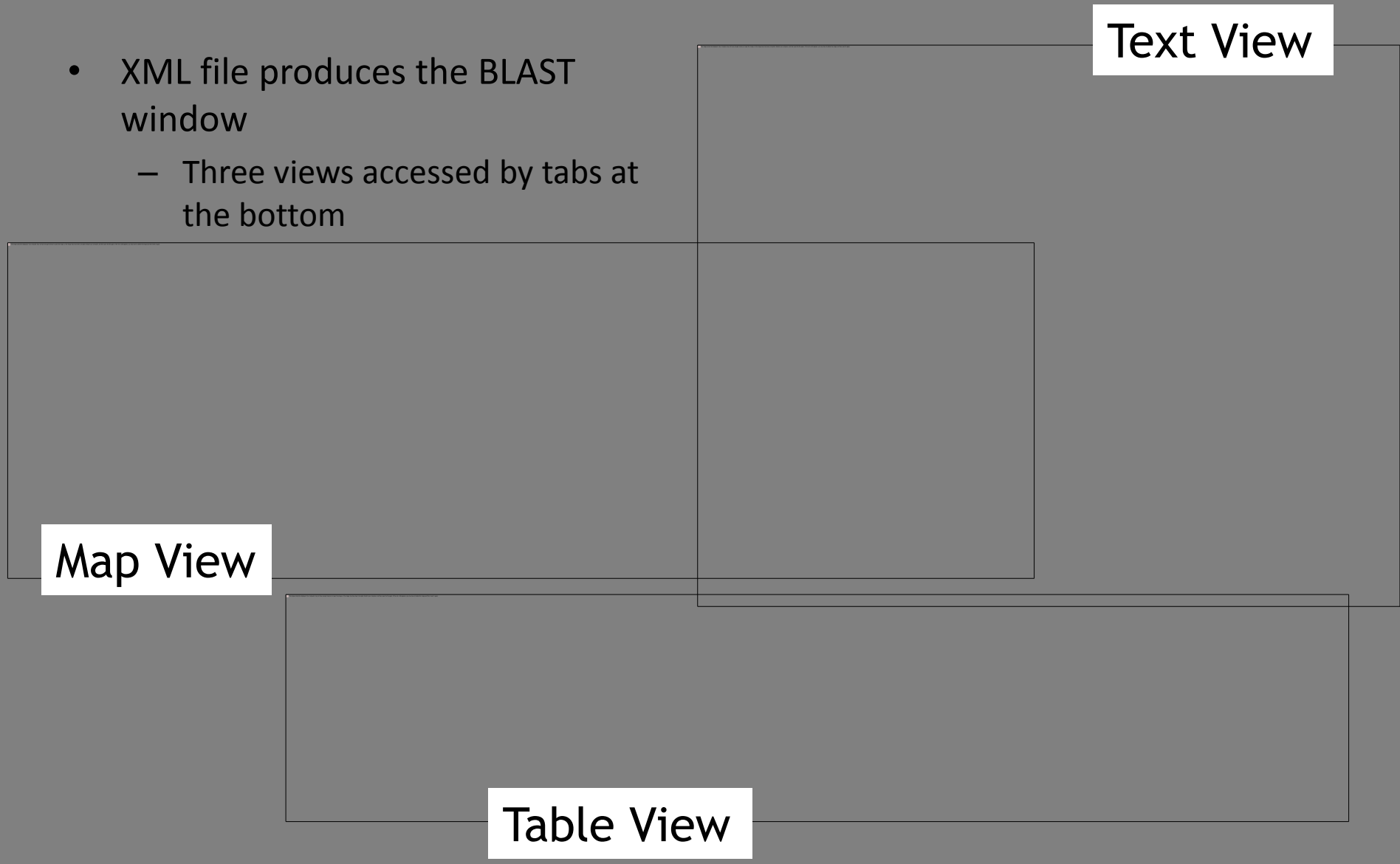


Viewing BLAST Results

- Double clicking in the Jobs Explorer returns the Report.htm file
- Complete results are found in the *.xml file
- All hits grouped in two files
  - PIR file contains the resulting alignment with the query
  - FASTA file has the unaligned sequences for only the hits

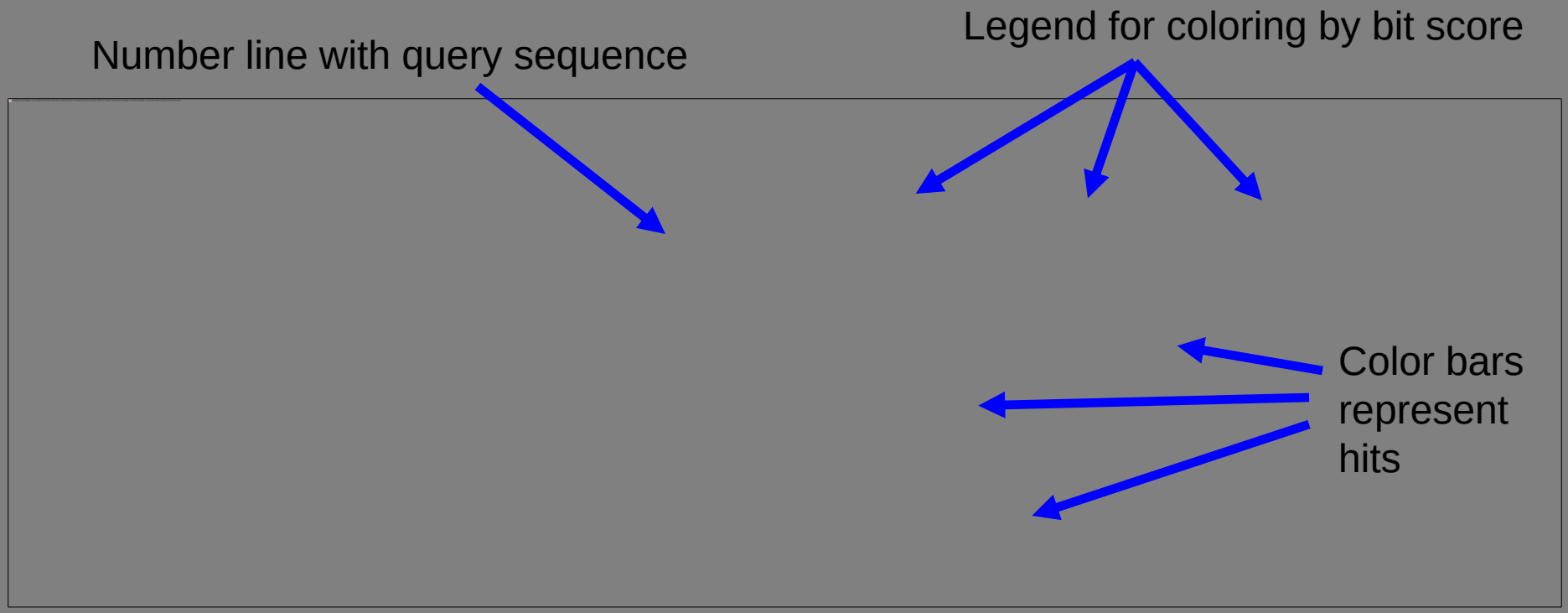
Viewing BLAST Results

- XML file produces the BLAST window
  - Three views accessed by tabs at the bottom



BLAST Results – Map View

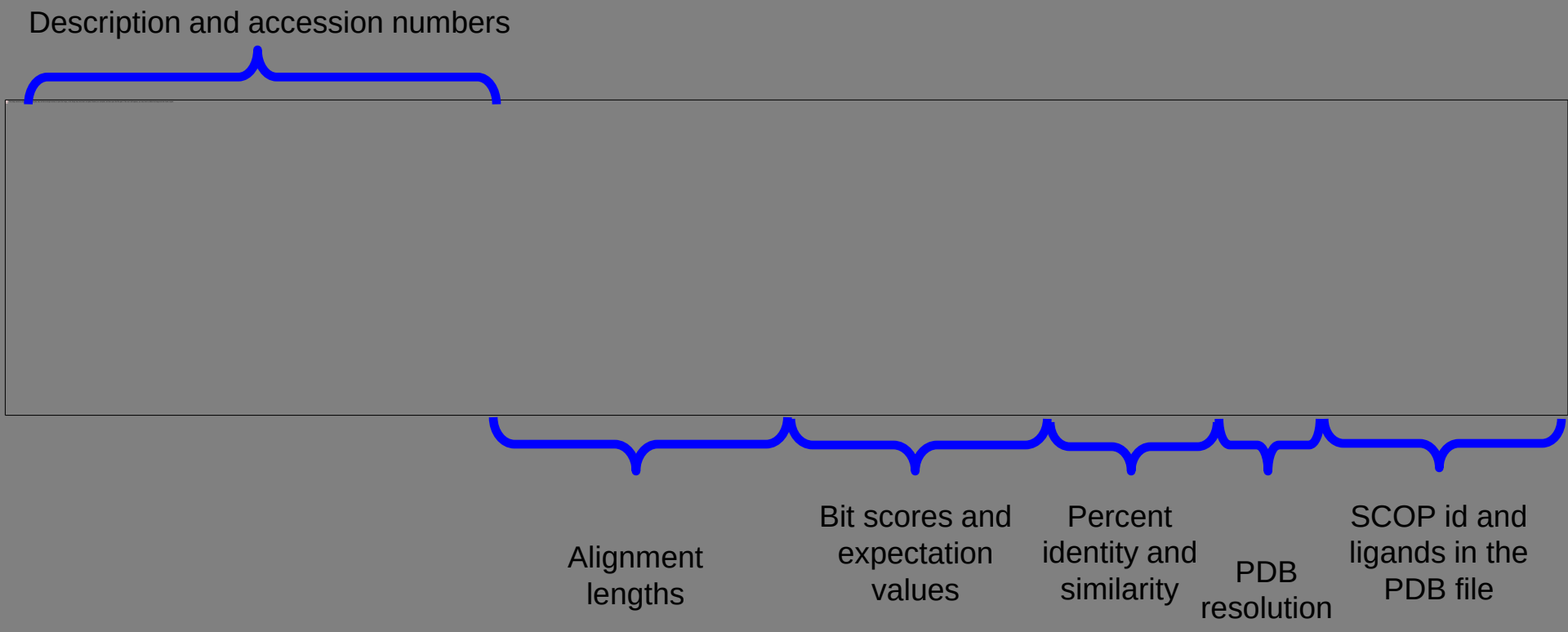
- Graphical representation of results



Mouse over the hit bars to give additional information

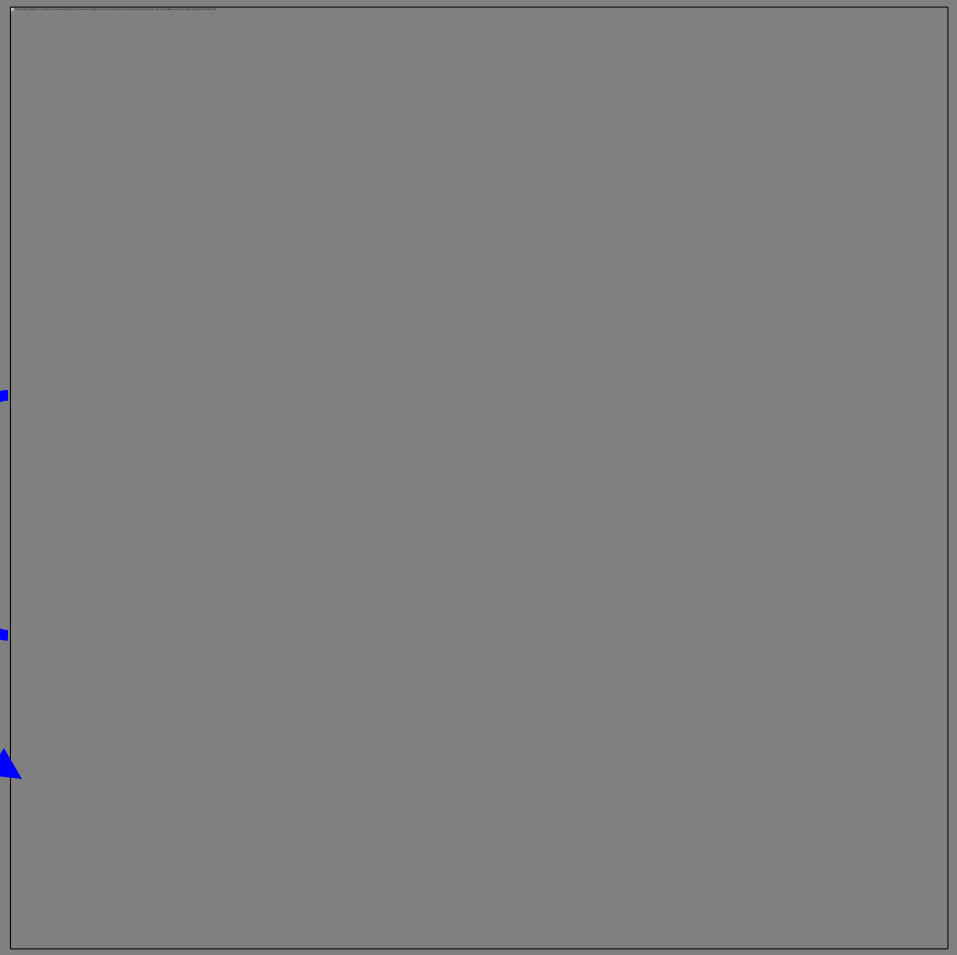
BLAST Results – Table View

- Tabulated data from the identified hits



BLAST Results – Text View

- The raw output from the BLAST program
- Contains
  - One line summaries of hits
  - Alignments
  - Additional information in the footer



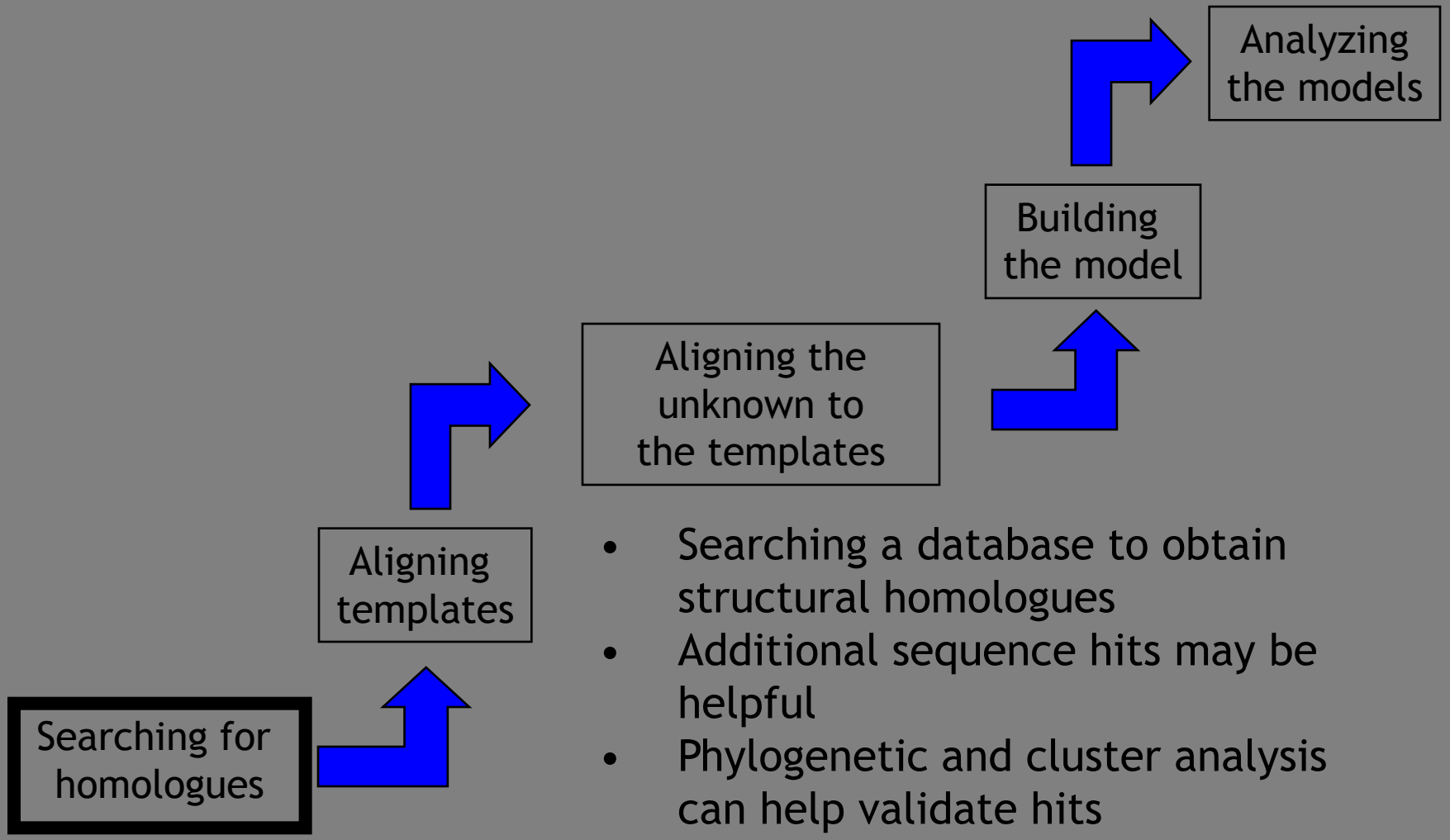
Selecting Hits from the Searches

- Want low expectation values
  - Good hits are in the range of 10^{-5} or lower for BLAST searches
  - Profile searches may not be that low
- Consider the length of the sequence alignment
  - Ideally want an alignment through entire unknown
  - At least cover the domain of interest in your unknown
- Consider residue match distribution
  - Residue matches throughout the length of the alignment
  - Not just alignment but actual matches
- Consider the alignment of conserved residues
  - Active site or binding site residues

Template Validation

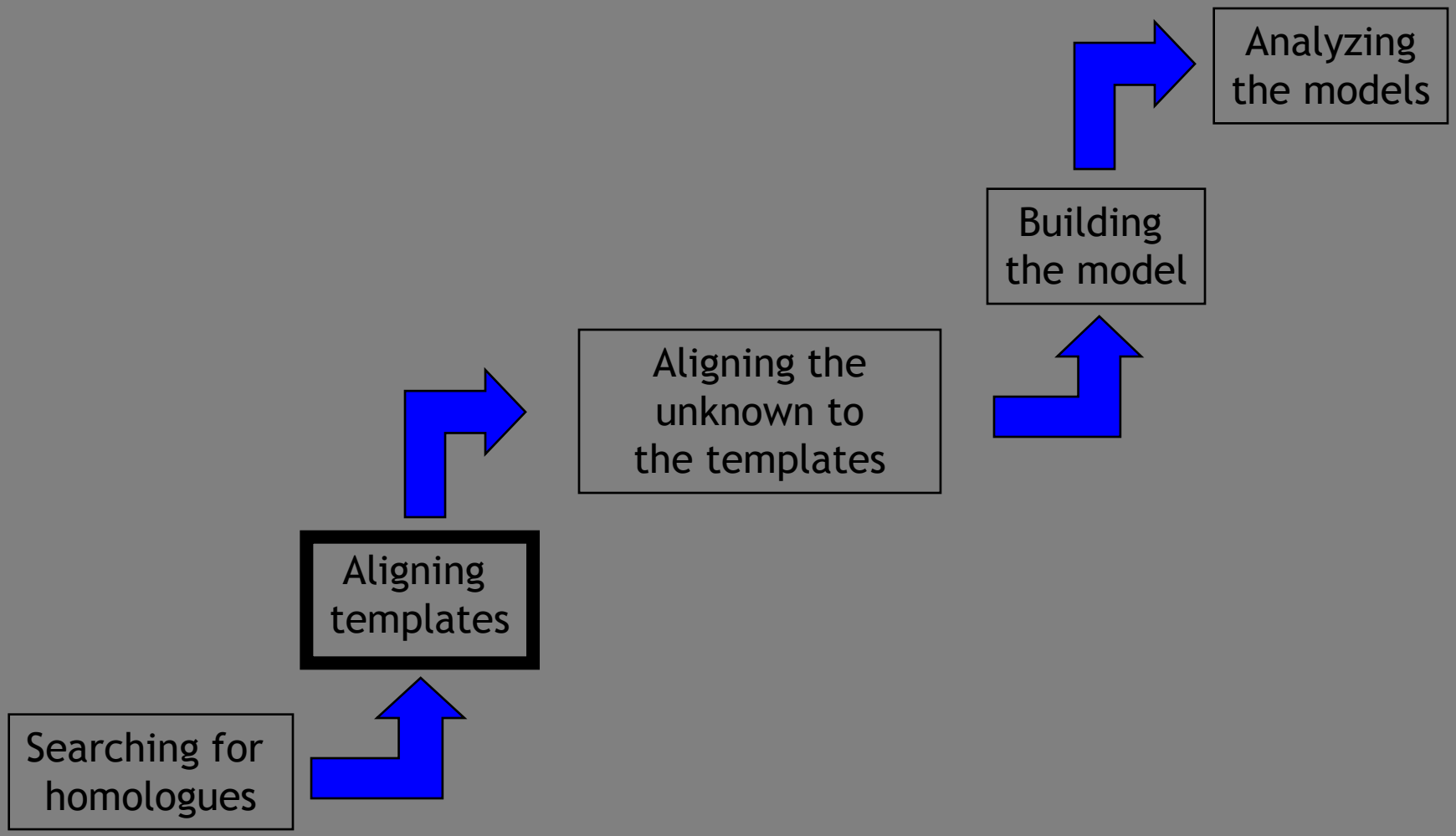
- Once you have hits for candidate templates, validate them
  - How similar are the candidates to each other and to the unknown?
  - Do you even have a structure available?
  - Examine the annotations for the PDB file
    - Same or related function as your unknown?
    - Similar organisms?
    - Active or inactive forms of protein?
    - Bound with a ligand or free?
  - For X-ray structures
    - What is the resolution?
    - Where are the missing residues?
  - Are the structures truly homologous in the region of interest?
    - Consider loop conformations

Structure Prediction by Homology Modeling



- Searching a database to obtain structural homologues
- Additional sequence hits may be helpful
- Phylogenetic and cluster analysis can help validate hits

Structure Prediction by Homology Modeling



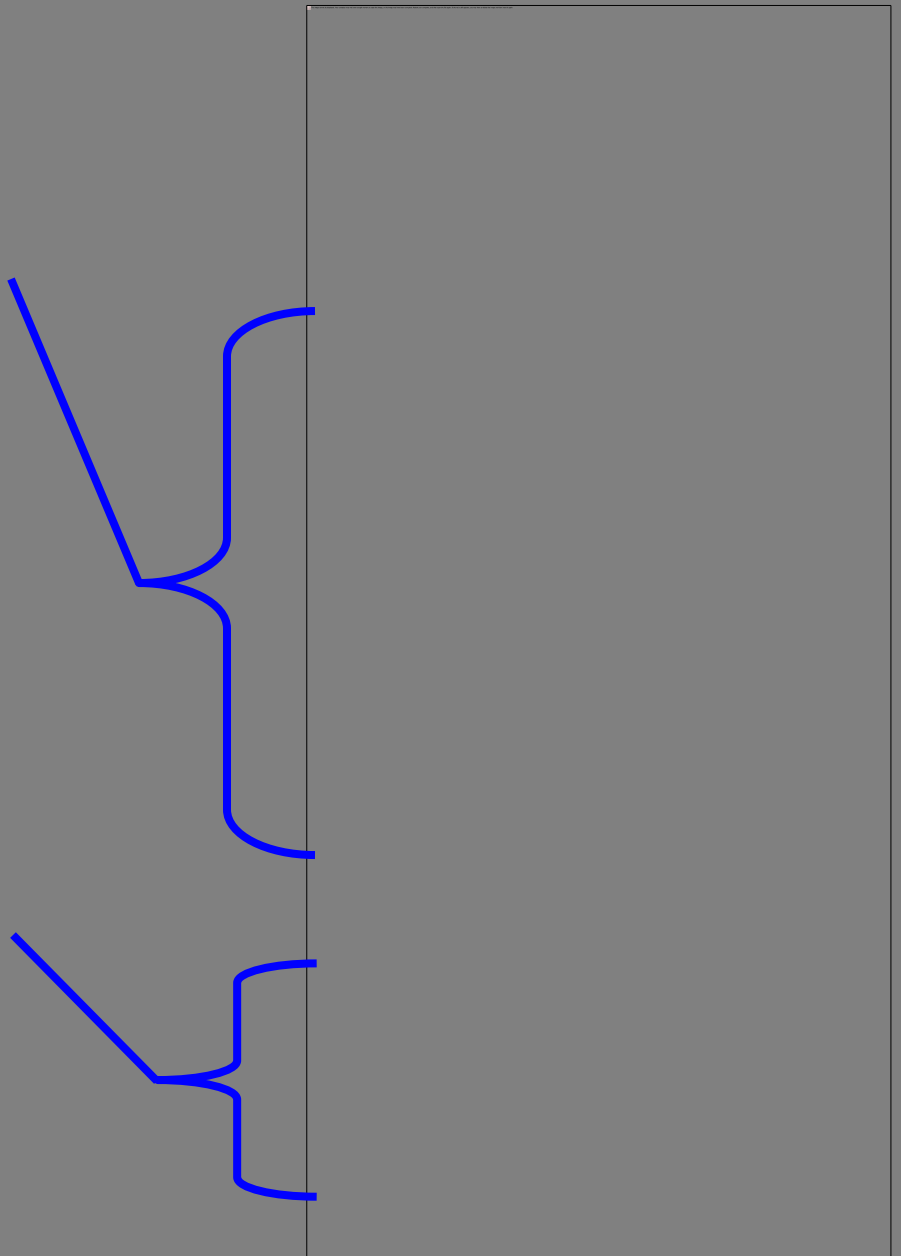
Once templates have been identified…

- The goal is to identify the structurally conserved regions
  - Structure is more conserved than sequence
- Within a protein family:
  - The inner core tends to be conserved
  - The α-helices and β-sheets tend to be orientated in the same way relative to one another
  - Sequences of the inner core regions are generally conserved

- Structural alignment
  - Find structural similarity among multiple templates
- Sequence alignment
  - The sequence alignment ideally should agree with the structural alignment
  - If only one template, could align sequence homologues to reinforce alignment of unknown to the one template
- Profile-profile alignment
  - Useful for aligning weak homologues
  - Unknown to templates if there is low sequence homology

Alignment Protocols

- Structure alignment protocols in Protein Modeling
  - Align and Superimpose Proteins
    - Uses Align123
  - Align Sequence to Templates
    - Uses BLAST, MODELER Align3D and Align123
  - Align Structures
    - Uses 3DMA
  - Align Structures (MODELER)
    - Uses MODELER Align3D
- Sequence alignment protocols in Sequence Analysis
  - Align Multiple Sequences
    - Runs Align123
  - Align Sequence Profiles
    - Runs SALIGN from MODELER



Align Structures Protocol

- Uses 3DMA
- Multiple alignment of protein structures
- Alignments are clustered to find related structures
- Only reports the residues that are aligned among all input proteins
- Use when aligning no more than two templates for homology modeling
- Results are reported back as a new sequence alignment and a superimposition of the structures based on the alignment

Align Structures (MODELER) Protocol

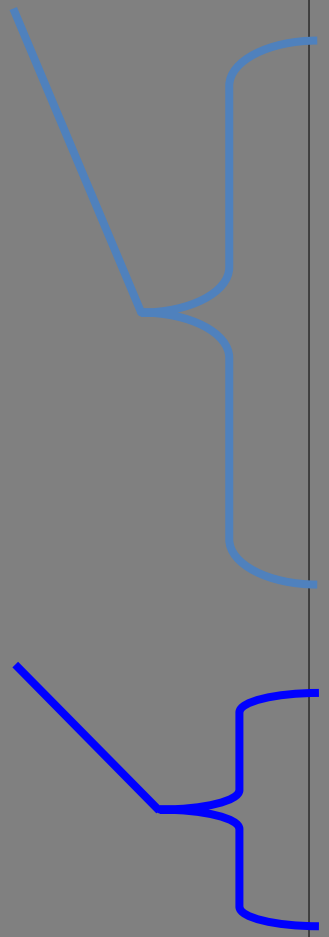
- Uses Align3D from MODELER
  - Aligns two or more template structures based on their structural similarity
  - Alignment is to a 3D framework rather than entire sequence or structure
  - Unlike **Align Structures**, this protocol:
    - Reports alignment even when the input structures are not homologous
    - Aligns residues even if all input proteins do not align in that region
  - Produces alignment suitable for homology modeling
  - Cannot align proteins that are drastically different in size
  - Also use to *Extend Existing Alignment*

Align Sequence to Templates Protocol

- Uses BLAST, MODELER Align3D and Align123
  - (Optional) Uses BLAST to search for homologous sequences to model sequence. Then creates sequence profile for aligning the model sequence to the templates
  - (Optional) Align template structures using MODELER Align3D
  - Align model sequence to templates. Depending on above settings and number of templates, alignment mode may be:
    - Multiple sequences
    - Sequence to Profile
    - Profile to Profile

Alignment Protocols

- Structure alignment protocols in Protein Modeling
  - Align and Superimpose Proteins
    - Uses Align123
  - Align Sequence to Templates
    - Uses BLAST, MODELER Align3D and Align123
  - Align Structures
    - Uses 3DMA
  - Align Structures (MODELER)
    - Uses MODELER Align3D
- Sequence alignment protocols in Sequence Analysis
  - Align Multiple Sequences
    - Runs Align123
  - Align Sequence Profiles
    - Runs SALIGN from MODELER



Align Multiple Sequences Protocol

- Uses Align123
- Multiple sequence alignment based on CLUSTALW
  - Takes original CLUSTALW score and adds a secondary structure score
- No tertiary structure is used
- Can consider secondary structure for the placement of gaps
  - Secondary structure must be identified first

Align Multiple Sequences Protocol

- Features
  - Individual weights assigned to each sequence in a partial alignment
    - Reduces effect of near duplicate sequences
  - Substitution matrices are varied according to the divergence of the sequences
  - Applies position-specific gap penalties
    - Encourages gaps in loops rather than secondary structures
  - Gaps appearing early in the alignment are given reduced gap penalties

Align Multiple Sequences Protocol

- Profiles of previously aligned sequences may be used
  - The profiles are simply prealigned sequences
  - Gaps are retained automatically
  - Must specify two separate sequence windows

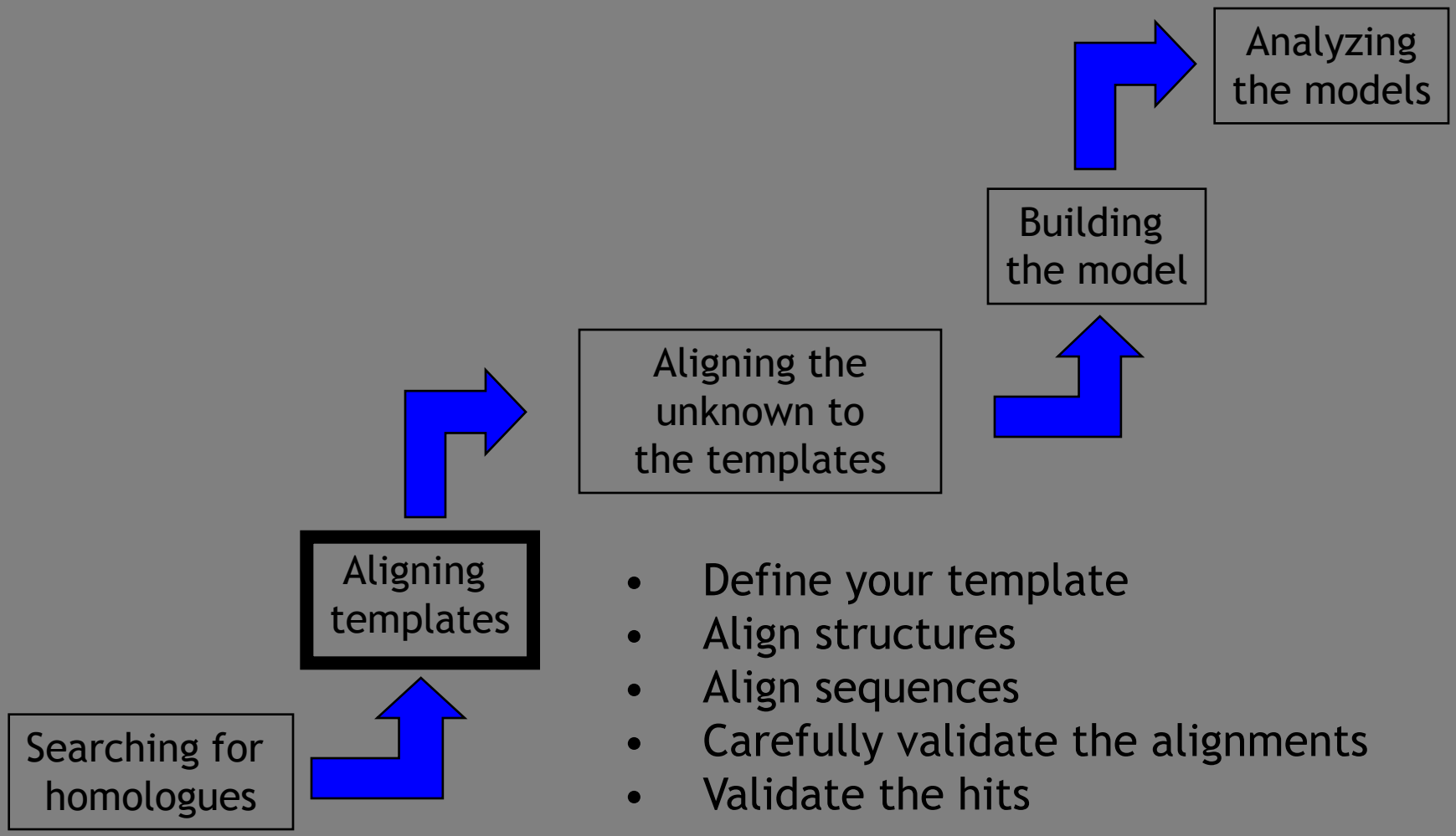
Aids to Alignment of Templates

- Use structural alignments
  - Structure is more conserved than sequence
  - Useful only when you have multiple templates
- Multiple sequence alignments
  - Vary scoring matrices and other parameters
- Use known experimental data
  - Site-directed mutagenesis
  - Cross-linking data
- Predicted secondary structure
  - Can be included in Align123 calculations
- Prediction of transmembrane helices
  - Could be very significant for transmembrane proteins such as GPCRs

Guidelines for Alignment of Templates

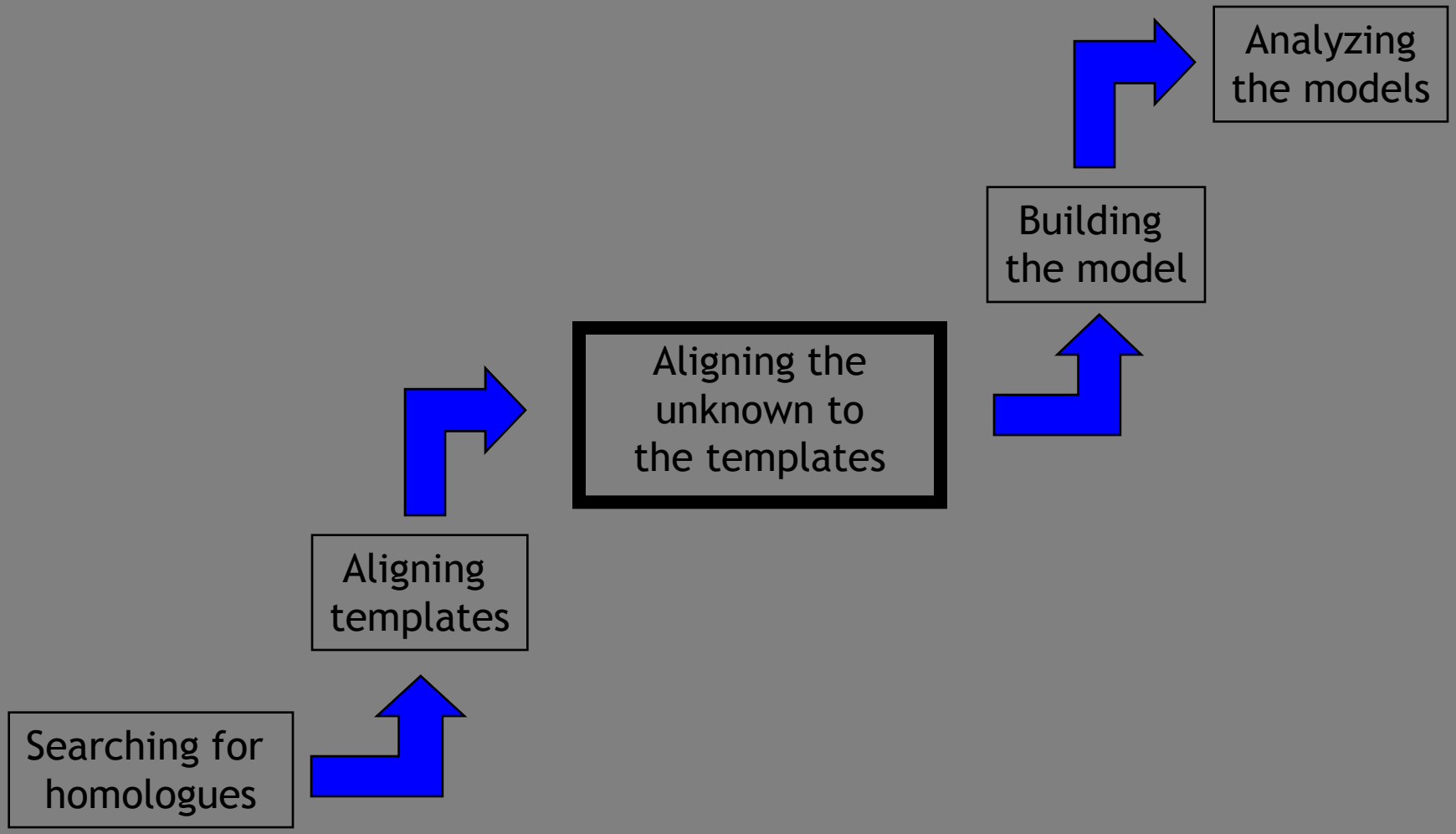
- Always validate the hits first
  - Are they valid homologues?
- Look for the similarities among the known or template structures
  - Compare with multiple structure and sequence alignments
  - Look at similarities and differences in both sequence and structure
- Look for key residues
  - Binding site or active site residues
  - Disulfide bridges
- Place gaps outside of regions of secondary structure
- Look out for Gly-nonGly or Pro-nonPro mutations
  - Often Gly/Pro residues have structural implications

Structure Prediction by Homology Modeling



- Define your template
- Align structures
- Align sequences
- Carefully validate the alignments
- Validate the hits

Structure Prediction by Homology Modeling

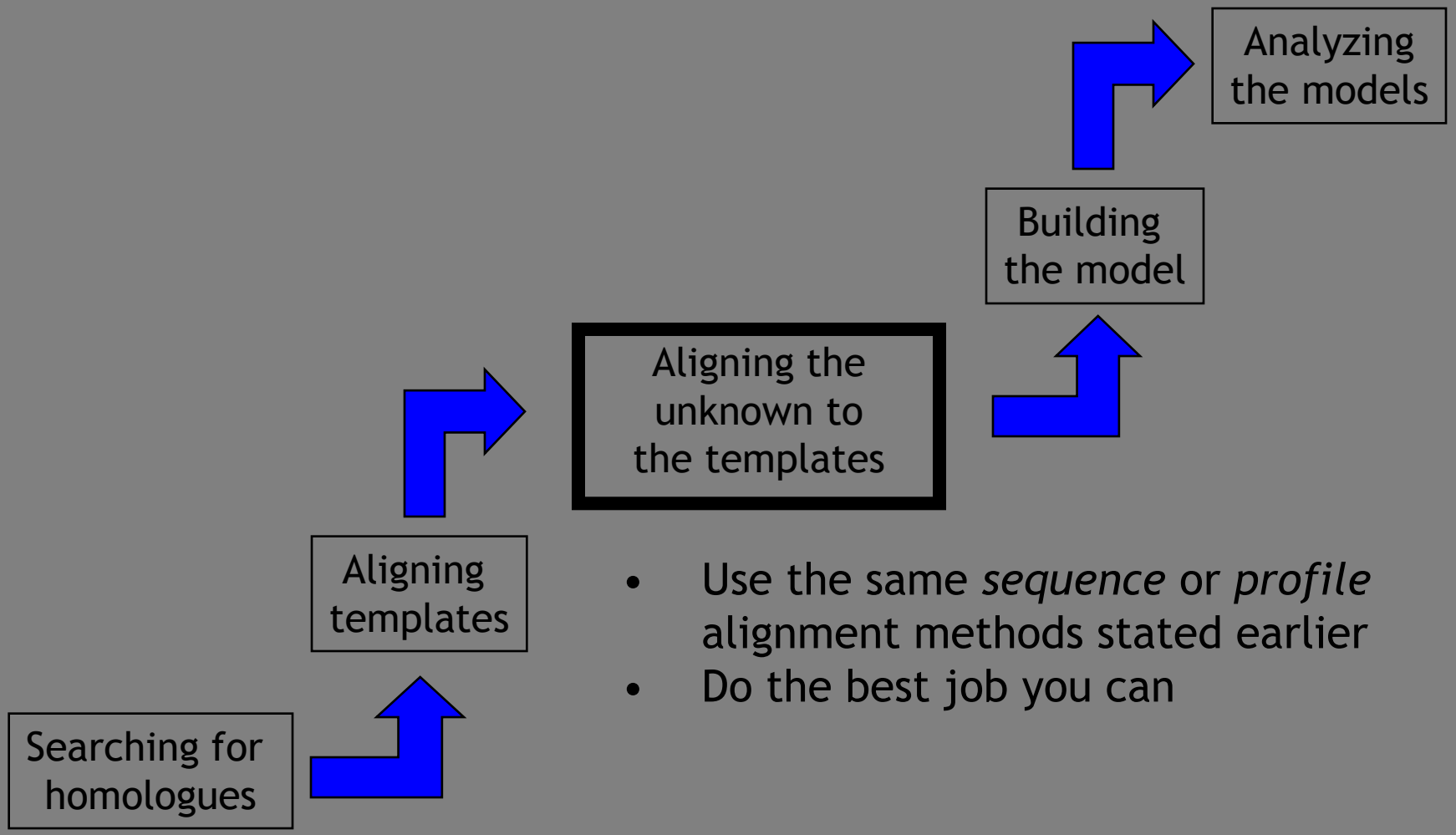


Alignment of Target to Template

- ***Sequence alignment is often the most critical step***
  - At 30% or higher sequence identity, the best CASP models have the most accurate alignments (Kryshtafovych *et al* (2005))

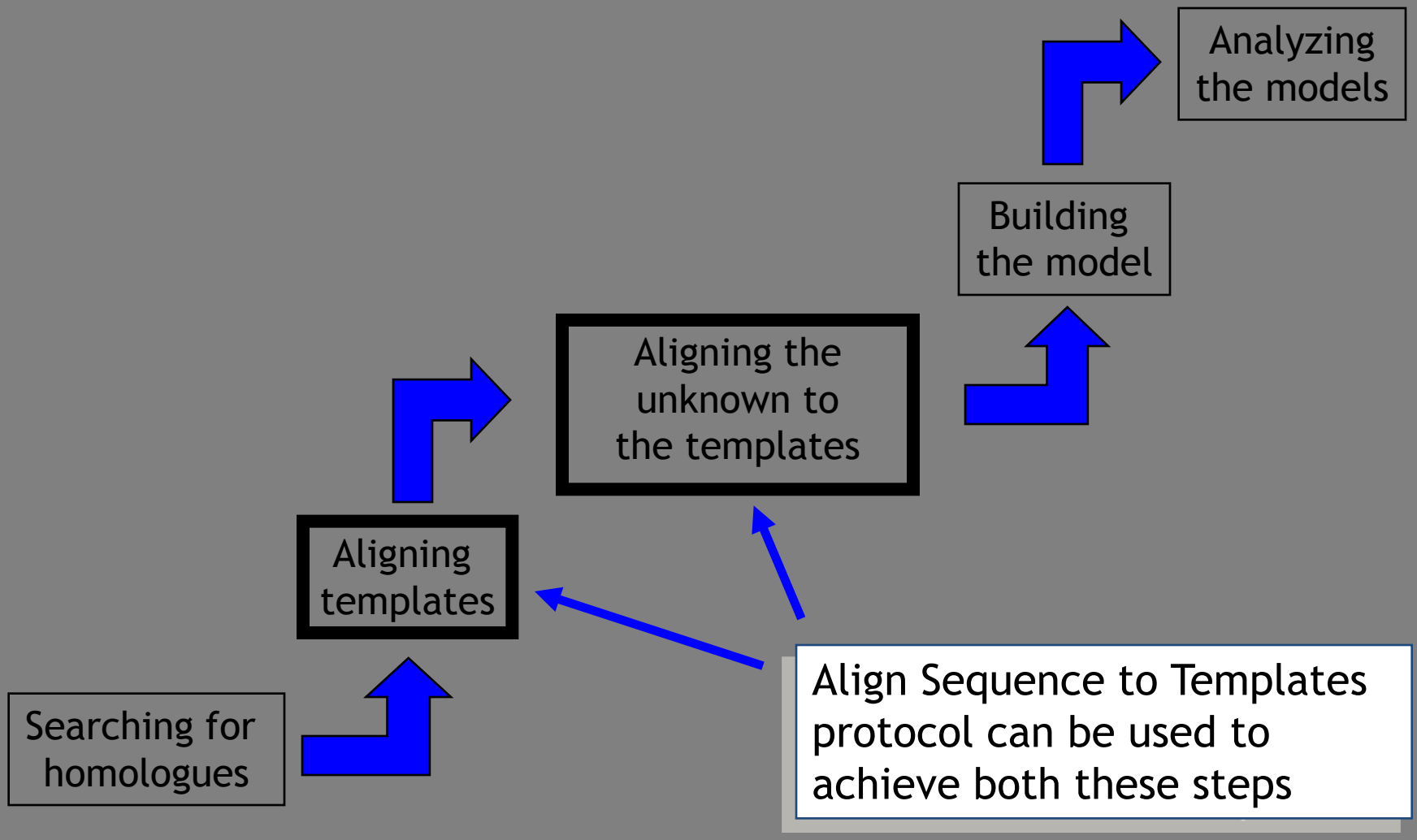
- Use structural alignments to verify sequence alignments of templates
- Use sequence alignments to fit target to templates
- Align unknown via sequence alignment protocols
  - Align Sequence to Templates
  - Align Multiple Sequences
  - Align Sequence Profiles
  - Manual methods

Structure Prediction by Homology Modeling

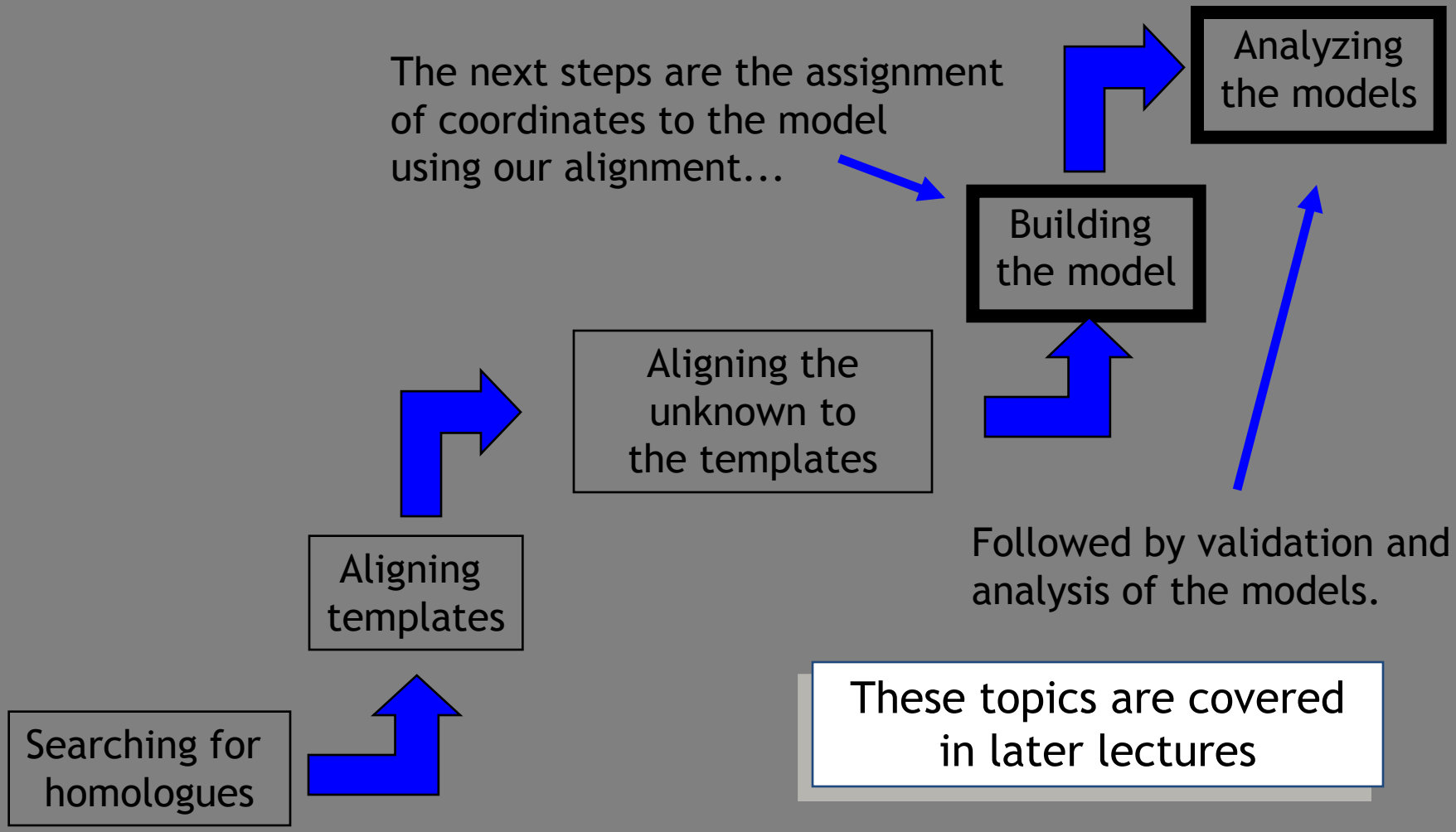


- Use the same *sequence* or *profile* alignment methods stated earlier
- Do the best job you can

Structure Prediction by Homology Modeling

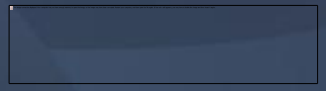


Structure Prediction by Homology Modeling

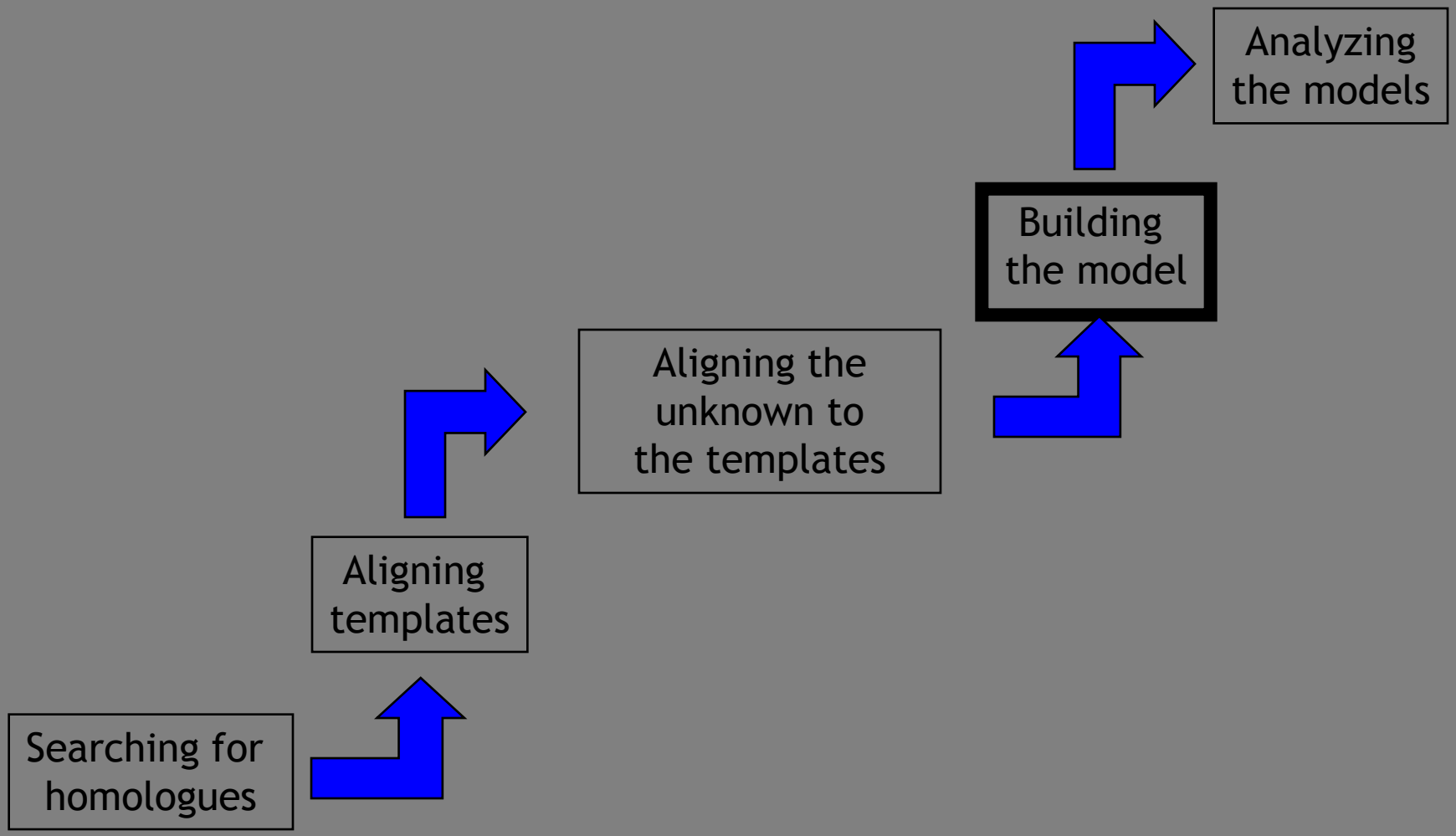


Automated Homology Modeling

Building the Structure



Structure Prediction by Homology Modeling



Homology Modeling with MODELER

- Automated homology modeling program
  - No manual assignment of coordinates
- Developed by Dr. Andrej Šali of U. California – San Francisco (UCSF)
- Requires at least one homologous or template structure
- Derives a 3D model close to the template structure while satisfying stereochemical and homology-derived restraints

MODELER is used in several protocols...

- From the Sequence Analysis for Homology Modeling Workshop:
  - Scan Sequence Profiles
    - Compares an input sequence profile against a sequence profile database
  - Align Sequence to Templates
    - Uses BLAST, MODELER Align3D and Align123
  - Align Structures (MODELER)
    - Uses MODELER Align3D
  - Align Sequence Profiles
    - Runs SALIGN from MODELER
- In this workshop...
  - Build Homology Models
    - Assigns coordinates to your model
  - Build Mutants
    - Mutates one or more residues and optimizes the local structure
  - Loop Refinement (MODELER)
    - Optimizes specified loop regions

General MODELER Protocol Requirements

- Usually requires:
  - An alignment between the model and the templates
  - The model sequence (the unknown)
  - The template structures (the references)
- Several options available
  - Number of models
  - Optimization level
  - Loop refinement
  - Additional restraints

Build Homology Models

Parameter Name	Parameter Value
⊟ Input Sequence Alignment	P41131
Input Model Sequence	P41131
Input Template Structures	1HX0,3DHP,1G94
Cut Overhangs	True
Disulfide Bridges	
Cis-Prolines	
Additional Restraints	
Copy Ligands	
Copy Chains	
⊞ Reference Template	
⊞ Number of Models	5
Optimization Level	High
⊞ Refine Loops	False

General Scheme

- Submit an alignment between the target sequence and at least one template sequence
  - Prepared using techniques discussed earlier
- Homology-derived restraints are calculated from the template(s)
  - For example: distances and dihedral angles
- Stereochemical restraints are added
  - For example: bond lengths, valence bond angles
- An objective function (F) is created
- Optimization of the objective function carried out in Cartesian space

Types of Restraints

- Stereochemical
  - Bond length, valence angles, dihedral angles, improper dihedral angles
  - Derived from CHARMM parameters
- Mainchain dihedrals angles
  - Main chain ϕ, ψ, and ω angles
- Sidechain dihedral angles
  - Sidechain χ_1, χ_2, χ_3, and χ_4 angles
- Distances
  - Mainchain C_α-C_α distances
  - Mainchain N-O distances
  - Sidechain-mainchain distances
  - Sidechain-sidechain distances
- Other special and user-defined restraints
  - Ligands
  - Additional disulfide bonds
  - NOE distances
  - Secondary structure
  - Symmetry

Description of Restraints

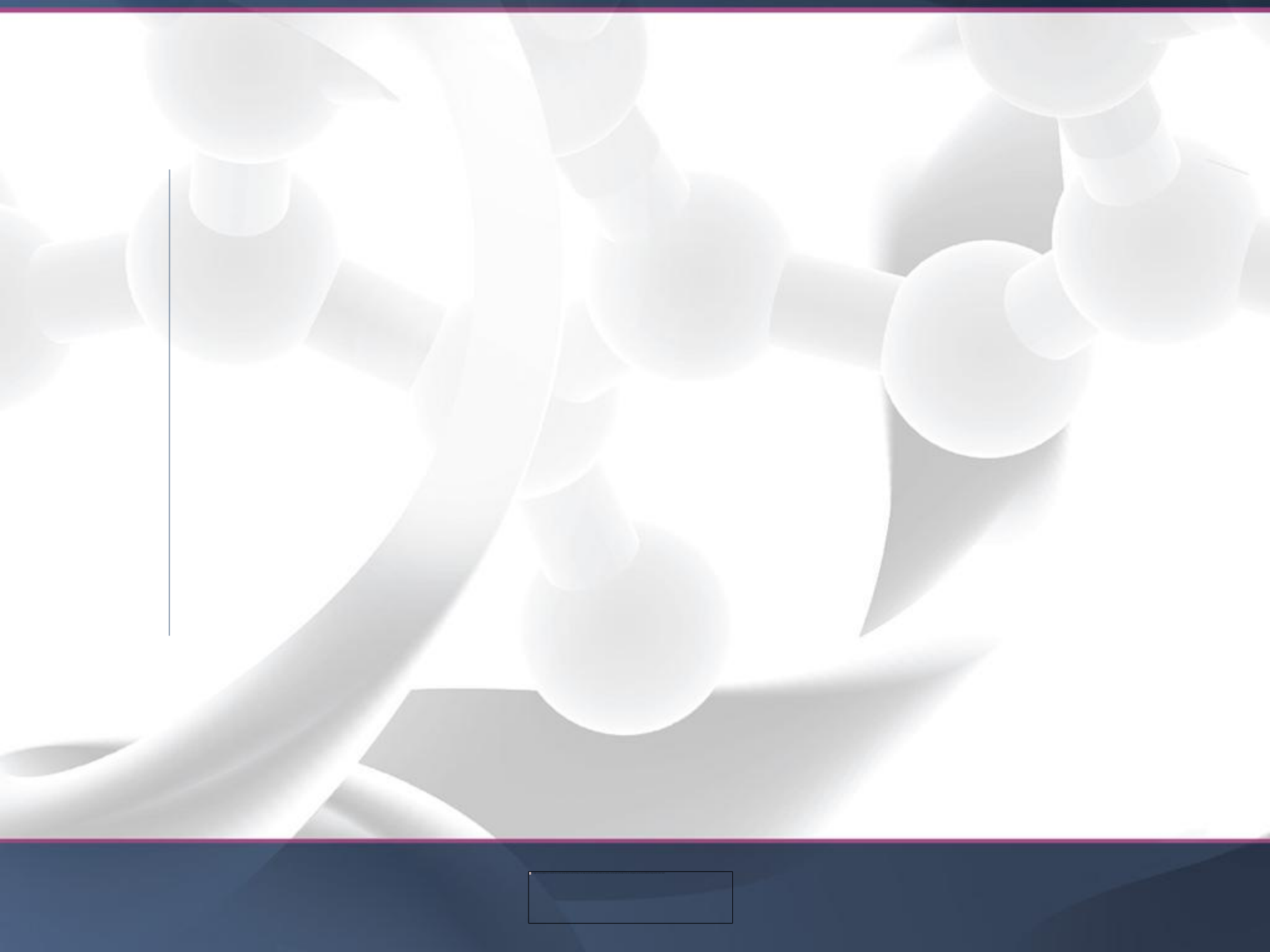
- Restraints are obtained empirically
- Taken from either:
  - Set of template structures
  - Derived database
- Derived database is part of MODELER and consists of:
  - 416 PDB entries
  - 1,233 sequence alignments
  - 78,495 residues
  - 230,396 pairs of equivalent positions
- From the derived database, correlations or relationships between restraints were obtained

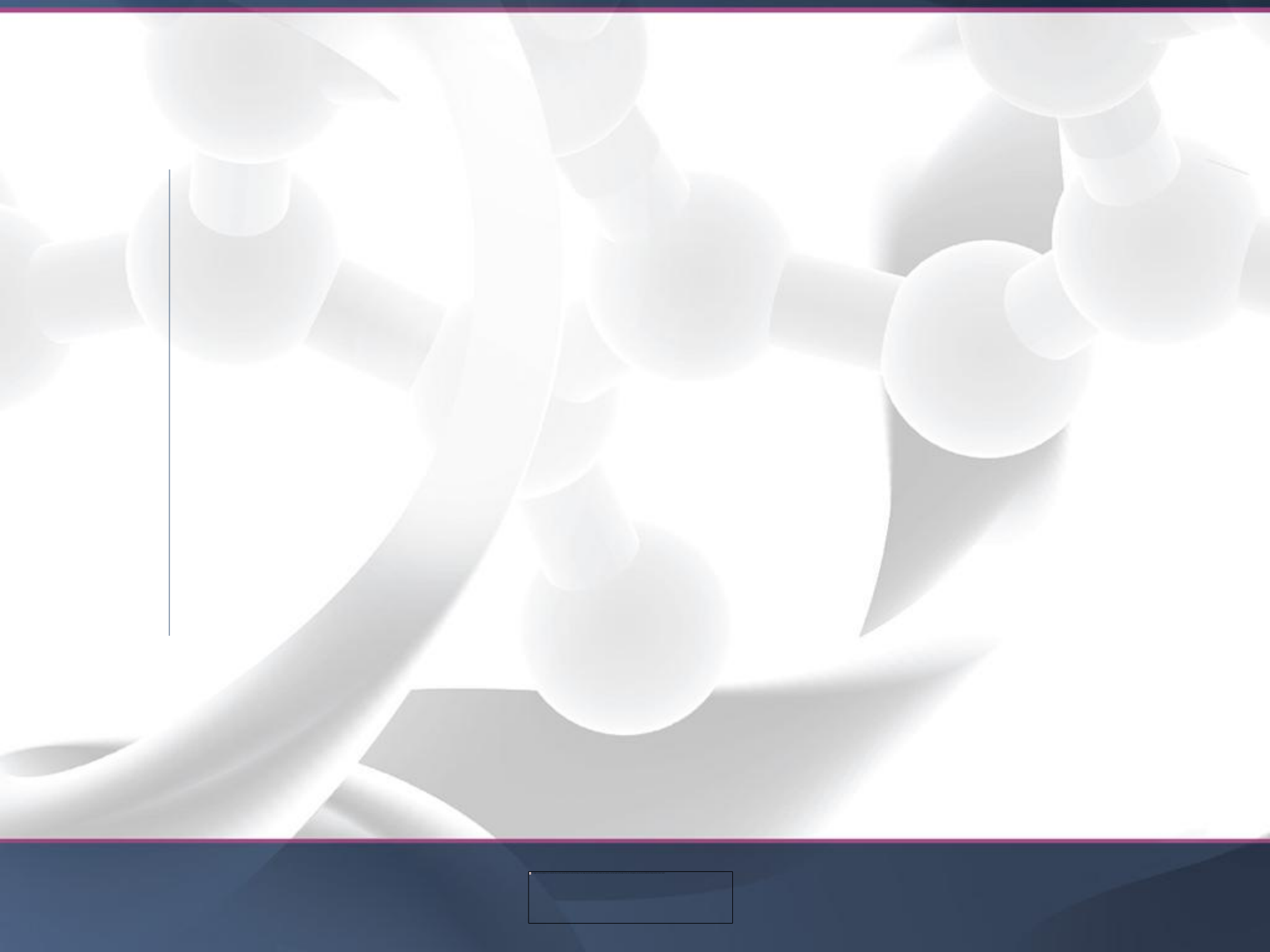
Basis Probability Density Functions

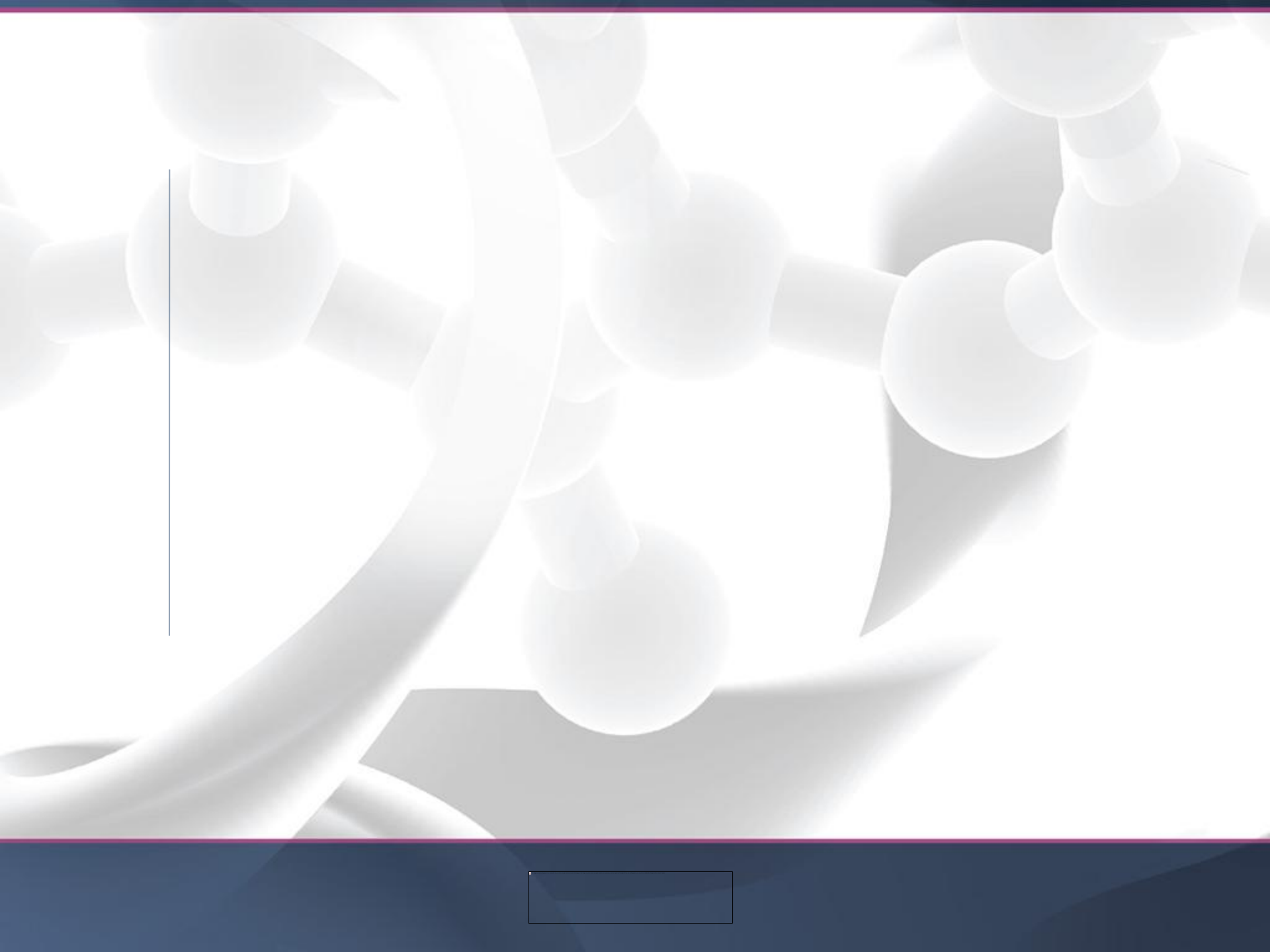
- Relationships are expressed as conditional probability density functions (PDF's)
  - Can be used directly as spatial restraints
- From the database:
  - A probability density function (PDF) is defined for each type of restraint
  - p(c) for a restraint c that is to be restrained.
- For example, given a mainchain dihedral angle, the PDF is derived from:
  - The type of a residue considered
  - Mainchain conformation of an equivalent residue
  - Sequence similarity between the two proteins

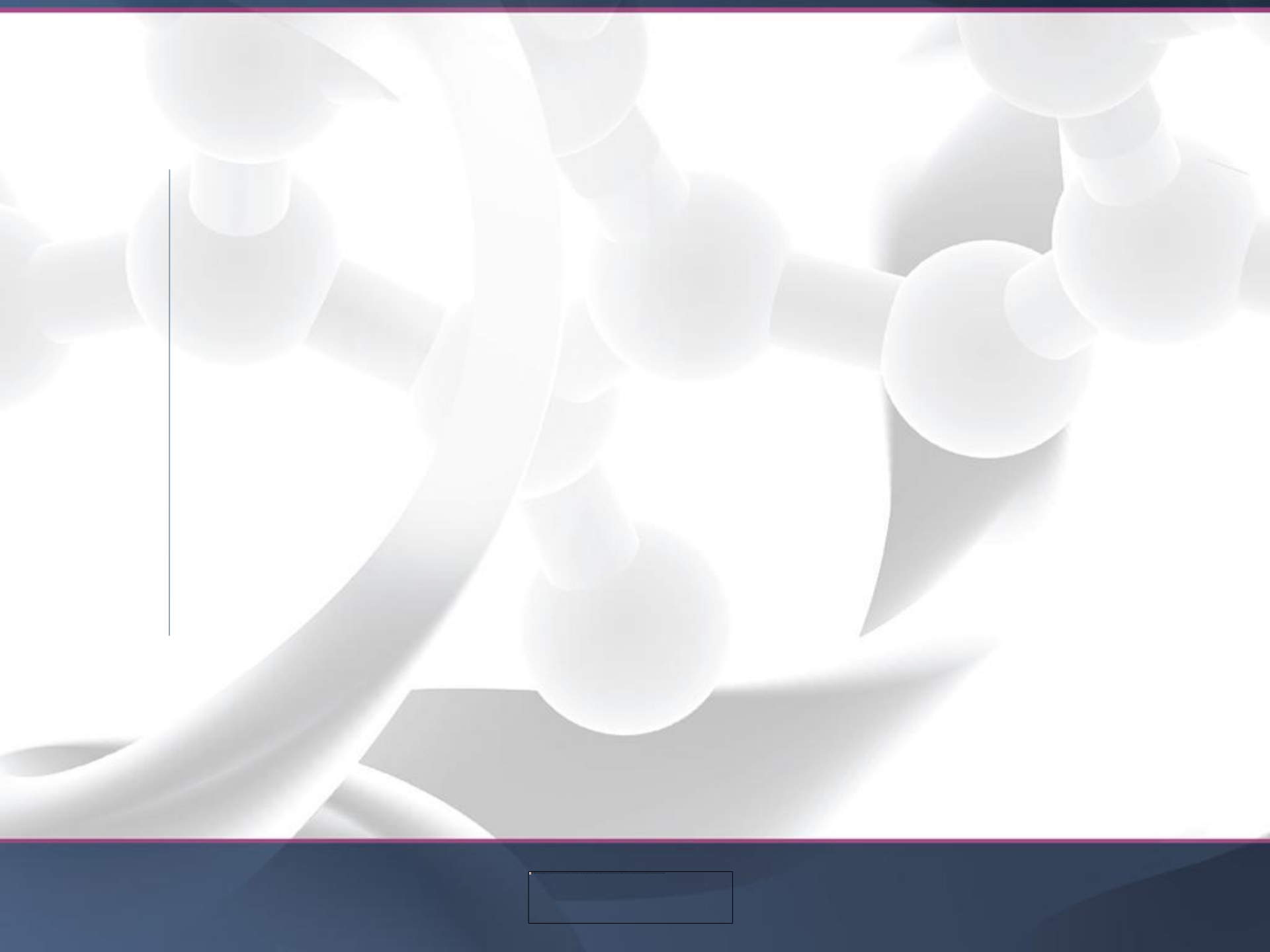
Bounds for Restraints

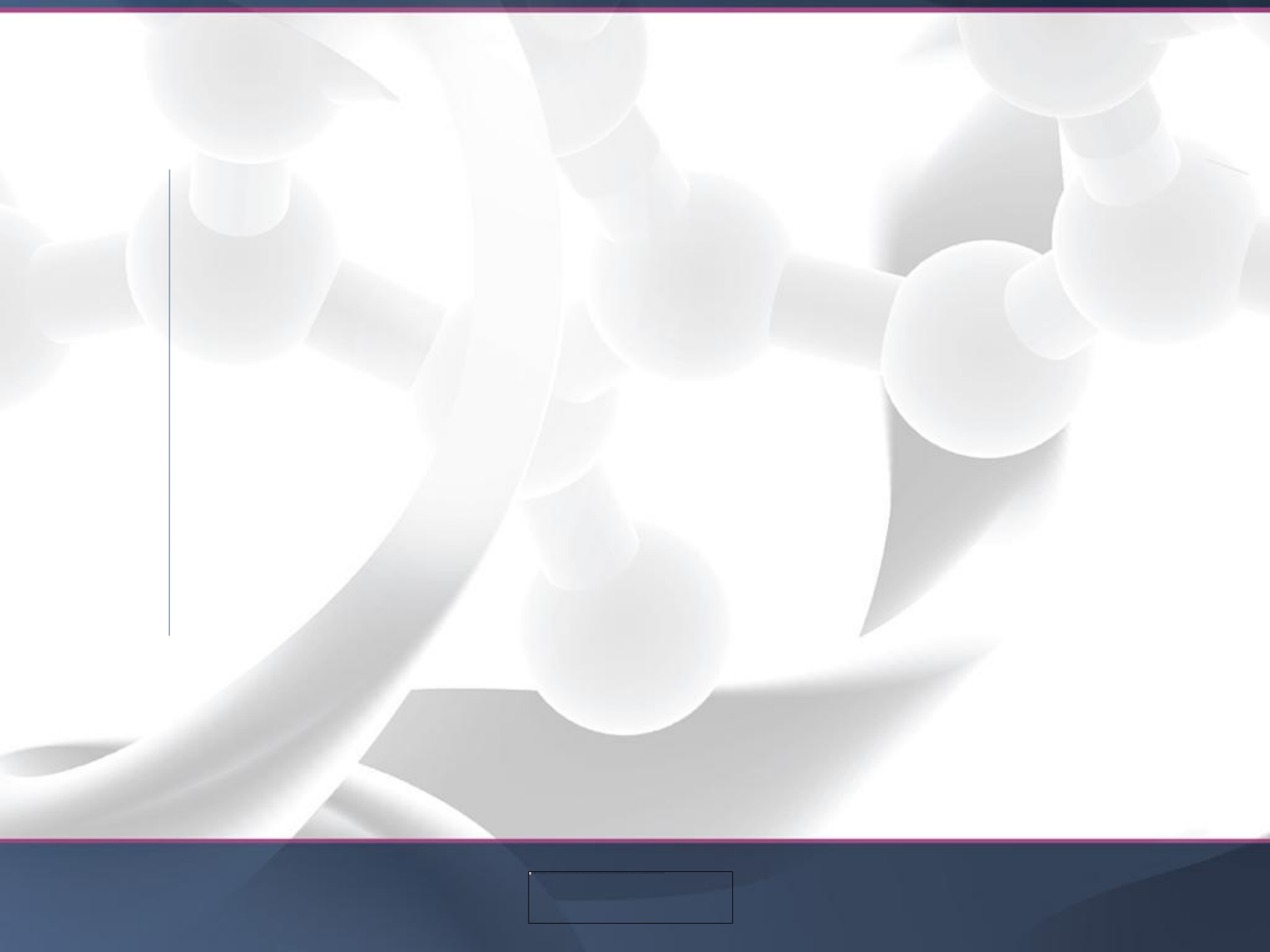
- PDF's allow for setting the upper and lower bounds for a given restraint based on either:
  - Template structures
  - Derived database
- Mathematical form for the restraints varies with the specific restraint
  - Gaussian, binormal, trigonometric, spline, and other functional forms are used
- If a feature is found outside the bounds
  - A pseudoenergy term and a restraining force are then applied











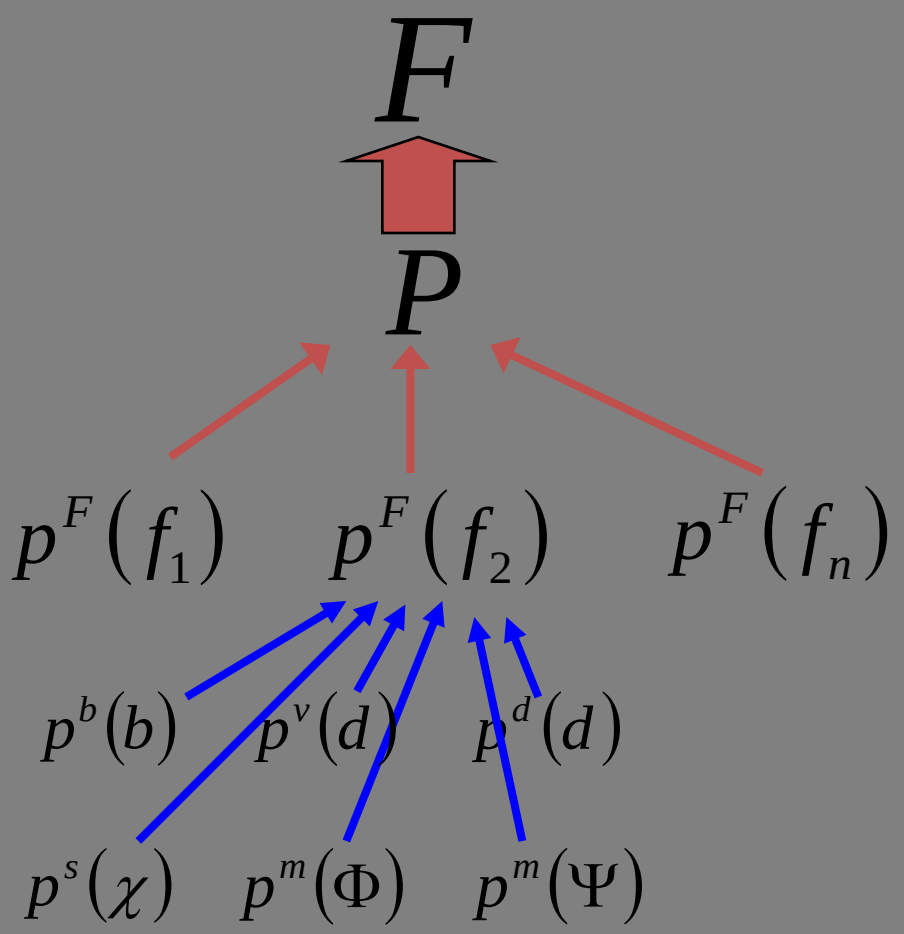
The Creation of the Objective Function

Objective function

Molecular PDF

Feature PDF's

Basis probability density functions



Structure Generation Using MODELER

- All templates are superimposed on the first template listed using the submitted alignment and the Cα positions
- Each atom in the modeled protein having an equivalent atom in at least one of the templates is assigned coordinates
  - Based on an average of all the template structures
- Undefined atoms are placed by using internal coordinates from a CHARMM residue topology library

Structure Generation Using MODELER

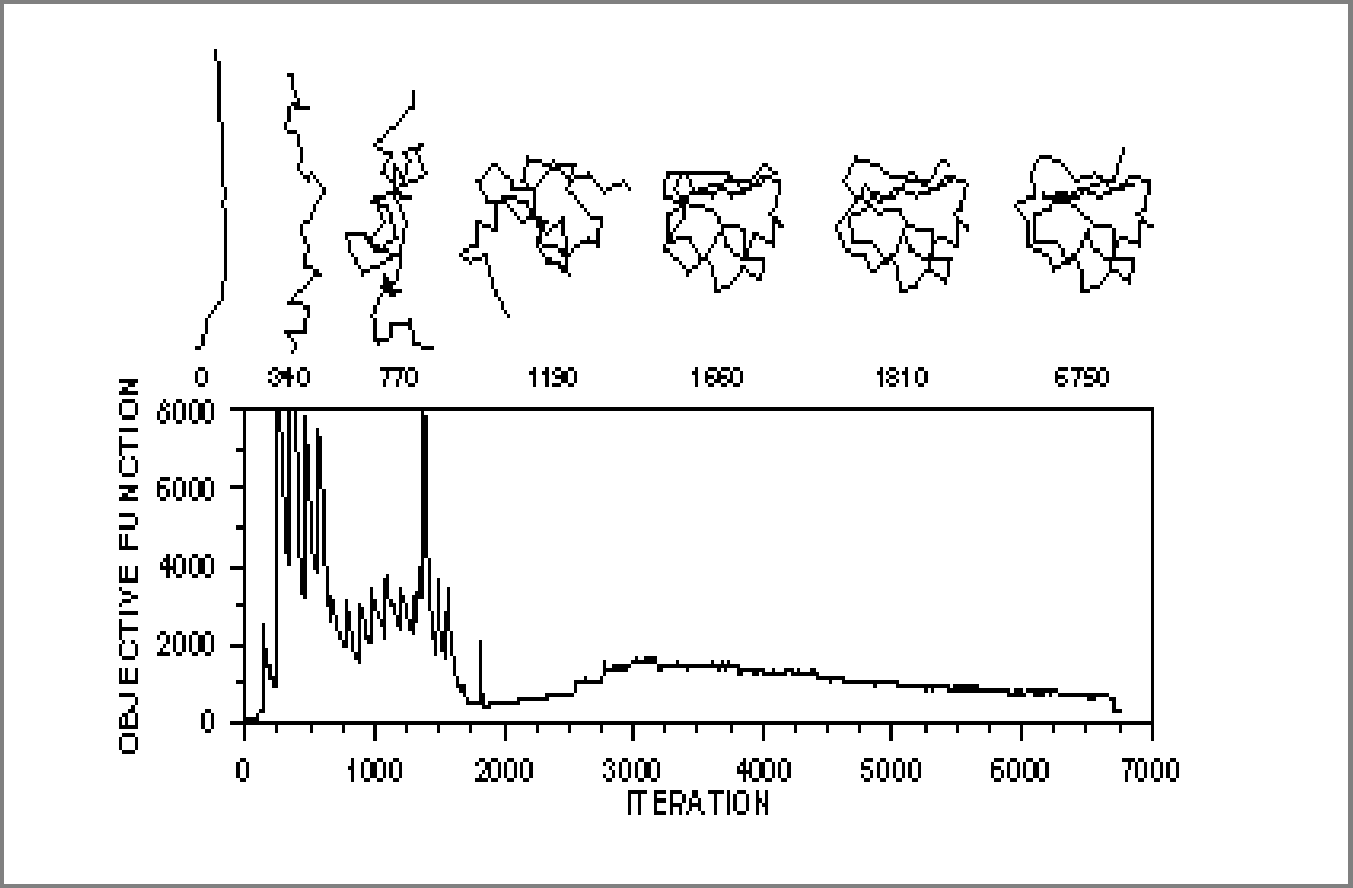
- The homology-based and stereochemical restraints are determined and the objective function is created
- A random coordinate shift of 4Å is added to each atomic position
  - By varying the shift, you can vary the models created
- Optimization begins using the variable target function method (VTFM)
  - Start with only local restraints
  - Add longer range restraints as the calculation proceeds
  - Eventually all restraints are being used and the full molecular PDF is being optimized

Structure Generation Using MODELER

- Further optimization carried out using restrained molecular dynamics and simulated annealing
- Within the Discovery Studio interface, four options for optimization:

High	VTFM minimization with thorough simulated annealing
Medium (the default)	VTFM minimization with medium simulated annealing
Low	VTFM minimization with very short simulated annealing
None	Coordinate assignment and VTFM minimization

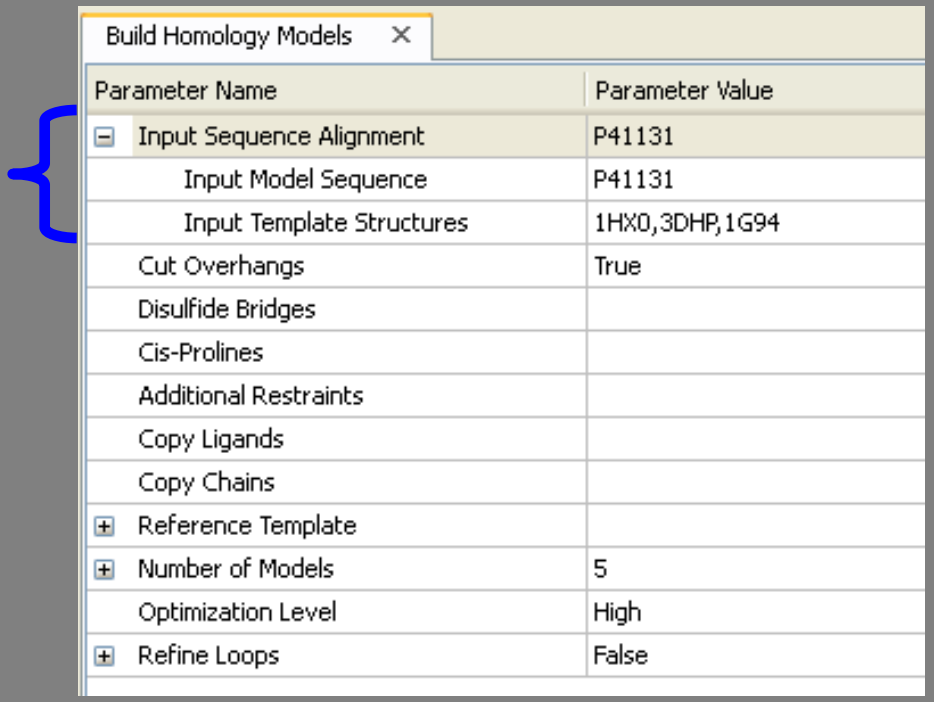
Structure Generation Using MODELER



VTFM minimization **Simulated annealing**

Build Homology Models Protocol

- Create homology models of a model sequence based on one or more template structures
- **Input Sequence Alignment**
  - Sequence window with desired alignment between templates and unknown
- **Input Model Sequence**
  - Name of unknown
  - Must be in the specified sequence window
- **Input Template Structures**
  - The templates or reference proteins
  - Sequences must be in the specified sequence window



Preparation of Template Structures

- PDB files may have problems
  - Missing atoms and residues
    - Not all parts of the molecule can be resolved especially hydrogen atoms
  - Non-standard residues
    - MODELER may not be able to model them
  - Presence of heteroatoms
    - Any non-polypeptide atom is a heteroatom which would include ligands
    - Must be handled separately in many cases
  - Presence of solvent
    - Listed as heteroatoms
  - Disorder in the structure
    - May result in multiple sets of atomic coordinates

Preparation of Template Structures

- Protein Health Tool panel
  - Used to identify poorly defined regions of the protein
- Protein Reports and Utilities Tools panel
  - Renumber sequences
  - Generate the Protein Report
  - Generate hydrophobicity plots
  - Split structures into distinct molecules
  - Clean protein molecules
    - Add missing atoms
    - Fix connectivity
    - Fix names
- Prepare Protein protocol (found under General Purpose folder)
  - Automatically cleans, adds and optimizes missing loops

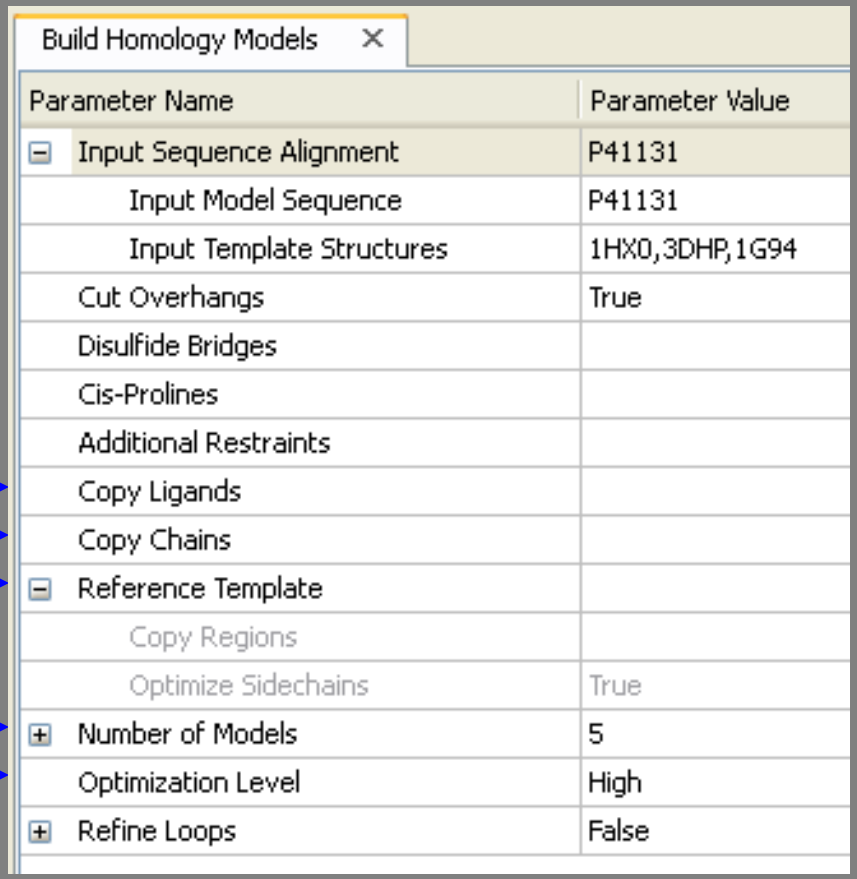
Prepare Protein

Used to:

- Patch missing atoms in a residue
- Optimize missing side chains
- Patch and optimize missing loops (optional)
  - Allow user to input the sequence (which will be used to patch missing loops)
  - Allow user to specify the disulfide bridges in loop
- Predict pK and optimize hydrogen position (optional)
- Keep selected water molecules from input (optional)
- Keep ligand from the input structure (optional)
- Detailed report of fixes applied

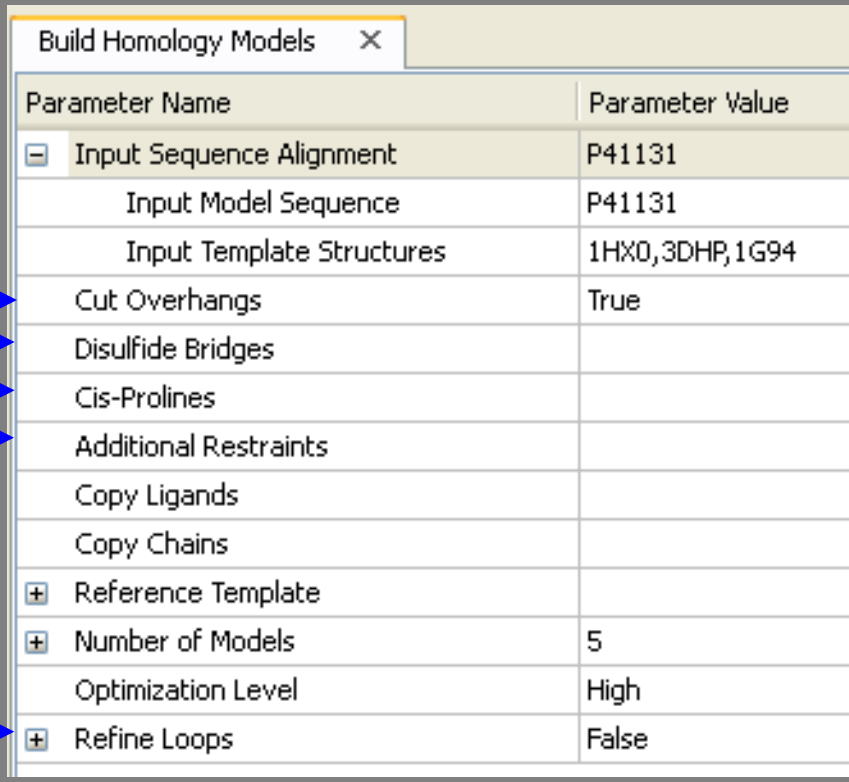
Build Homology Models Protocol

- **Copy Ligands**
  - Specifies a list of ligands to be transferred to the model
- **Copy chains**
  - Specifies a list of chains (not found in the alignment) to be transferred to the model
- **Reference Template**
  - Regions in the template which are copied by not optimized
- **Number of Models**
  - How many models are to be built?
  - Same alignment each time
  - But different random scattering of atoms
  - More models, the longer the run
- **Optimization Level**
  - As described earlier



Build Homology Models Protocol

- **Cut Overhangs**
  - Clip N- and C-termini if no matching template
- **Disulfide Bridges**
  - Set cysteine pairs that form disulfide bonds
  - Useful when a known disulfide bridge does not occur in template
- **Cis-Prolines**
  - Specify individual proline residues
- **Additional Restraints**
  - Add distance, secondary structure, or symmetry restraints
- **Refine Loops**
  - Further optimization of loop regions



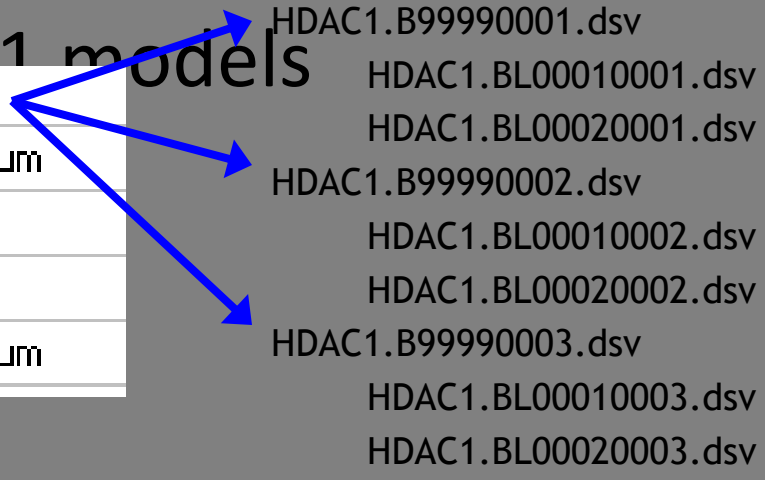
Loop Modeling

- Can be accomplished in two places
- While building the models with the **Build Homology Models** protocol
  - Use **Refine Loops** parameter
  - Creates multiple models with loop variations
  - No control over loops to refine
- Post-processing of models
  - **Loop Refinement (MODELER)** protocol
    - Same algorithm as **Build Homology Models** | **Refine Loops**
    - User specifies what loops to refine
  - **Loop Refinement** protocol
    - CHARMm-based method named LOOPER
    - User specifies what loops to refine
    - Can include implicit membrane

Loop Modeling and Refinement

- In the Build Homology Models protocol...
- Creates additional models as loops are refined
  - **Number of Models** parameter specifies the number of basic models created
  - **Refine Loops | Number of Models** creates additional models from each basic model
- For example, with HDAC1 models

⊞ Number of Models	3
Optimization Level	Medium
⊟ Refine Loops	True
Number of Models	2
Optimization Level	Medium



MODELER Loop Generation

- Loop refinement uses a different energy function from that used for the entire model
- Based on:
  - Stereochemical features from CHARMM residue topology definitions
  - Main chain and side chain dihedrals from basis PDF's
  - Nonbonded and solvation energies determined from a potential of mean force
- Effective up to 10-12 residues

Loop Refinement Protocol

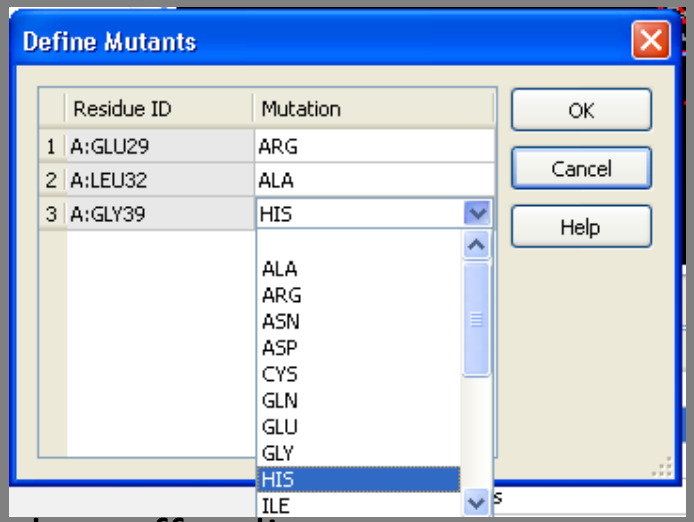
- Alternative protocol using the LOOPER algorithm
  - Force field based method using CHARMm
  - Requires properly atom typed protein
- Optimizes the structure of a specified loop region
  - Non-terminal loop region
  - Gen
    - M
- User s

Loop Refinement

Parameter Name	Parameter Value
Input Typed Protein Molecule	HDAC1.B99990003:HDAC1.B99990003
Loop	:TYR16-:ARG26
Advanced	
Maximum Number of Models	50
Maximum Number of Models to Save	50

Build Mutants

- Structure of “wild-type” is known
- Model the structure of mutants
- Uses Modeler
- Optimizes the conformation of:
  - mutated residues
  - surrounding residues within a specified cutoff radius



Build Mutants

Parameter Name	Parameter Value
Input Protein Molecule	1HX0:1HX0
Mutants	A:GLU29ARG,A:LEU32ALA,A:GLY39HIS
Disulfide Bridges	
Cis-Prolines	
Additional Restraints	
⊞ Number of Models	5
Optimization Level	High
Cut Radius	4.5

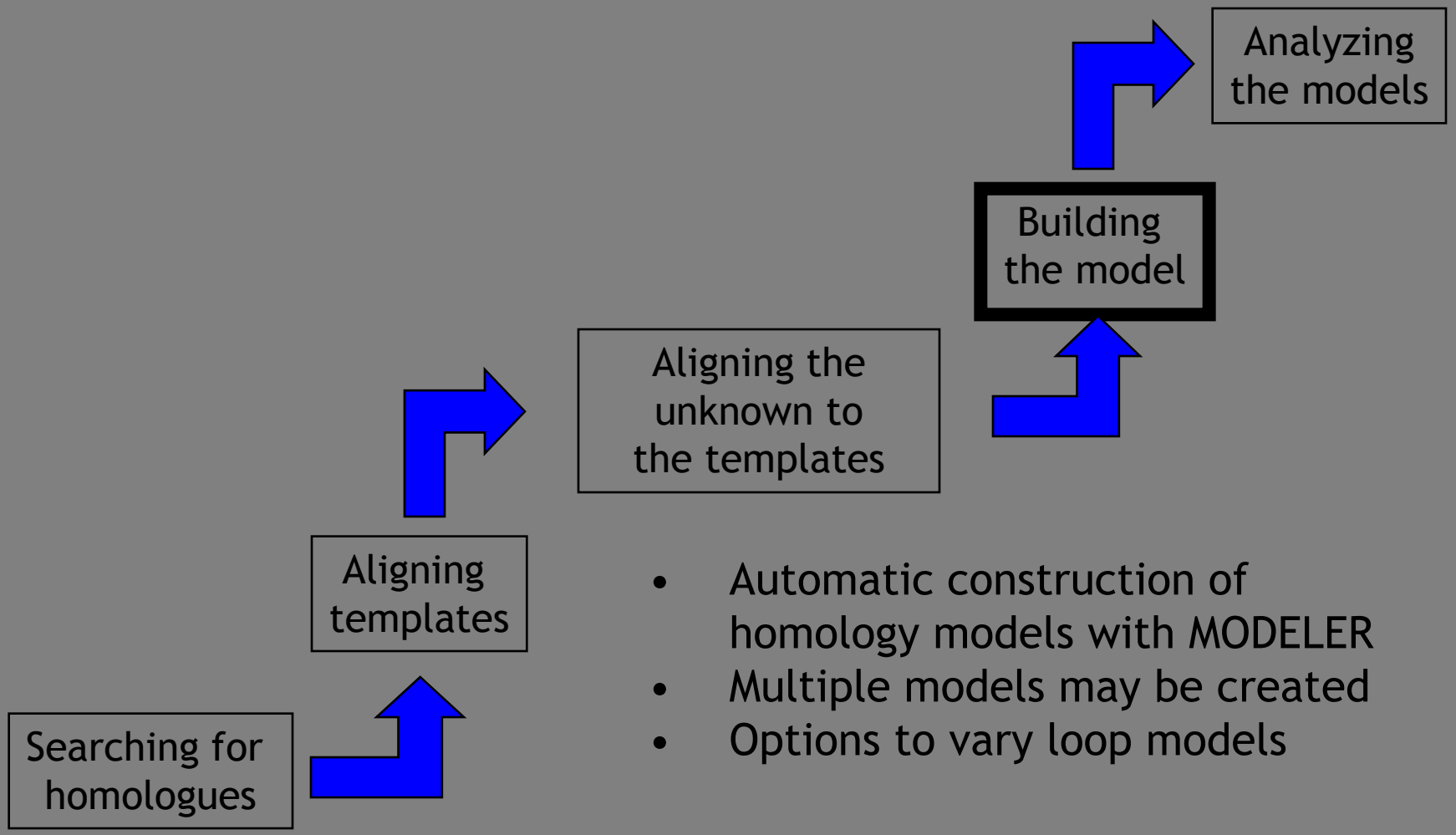
Modeling of Multiple Domains

- Generally, alignments are made to individual domains
  - One folding domain
- But with multiple domains...
  - There must be knowledge of the 3D orientation of the domains to each other
  - MODELER requires spatial restraints between the domains to place them
  - If the information is not available from templates, MODELER will not be able to place the individual domains
- Goal is to orient domains relative to each other in the same reference frame

Modeling of Multiple Domains

- Four scenarios
  - All domains are in a continuous chain in the template
    - Simply build model as before once an accurate alignment is attained
  - Domains from different templates that can be superimposed
    - Only use the appropriate parts of the templates to build the model
  - Model domains separately but can use a complete reference structure to orient modeled domains
    - Superimpose individual modeled domains onto reference structure
  - Model domains separately then use protein-protein docking to position domains
    - Used when no or limited information is available about domain orientations

Structure Prediction by Homology Modeling



- Automatic construction of homology models with MODELER
- Multiple models may be created
- Options to vary loop models

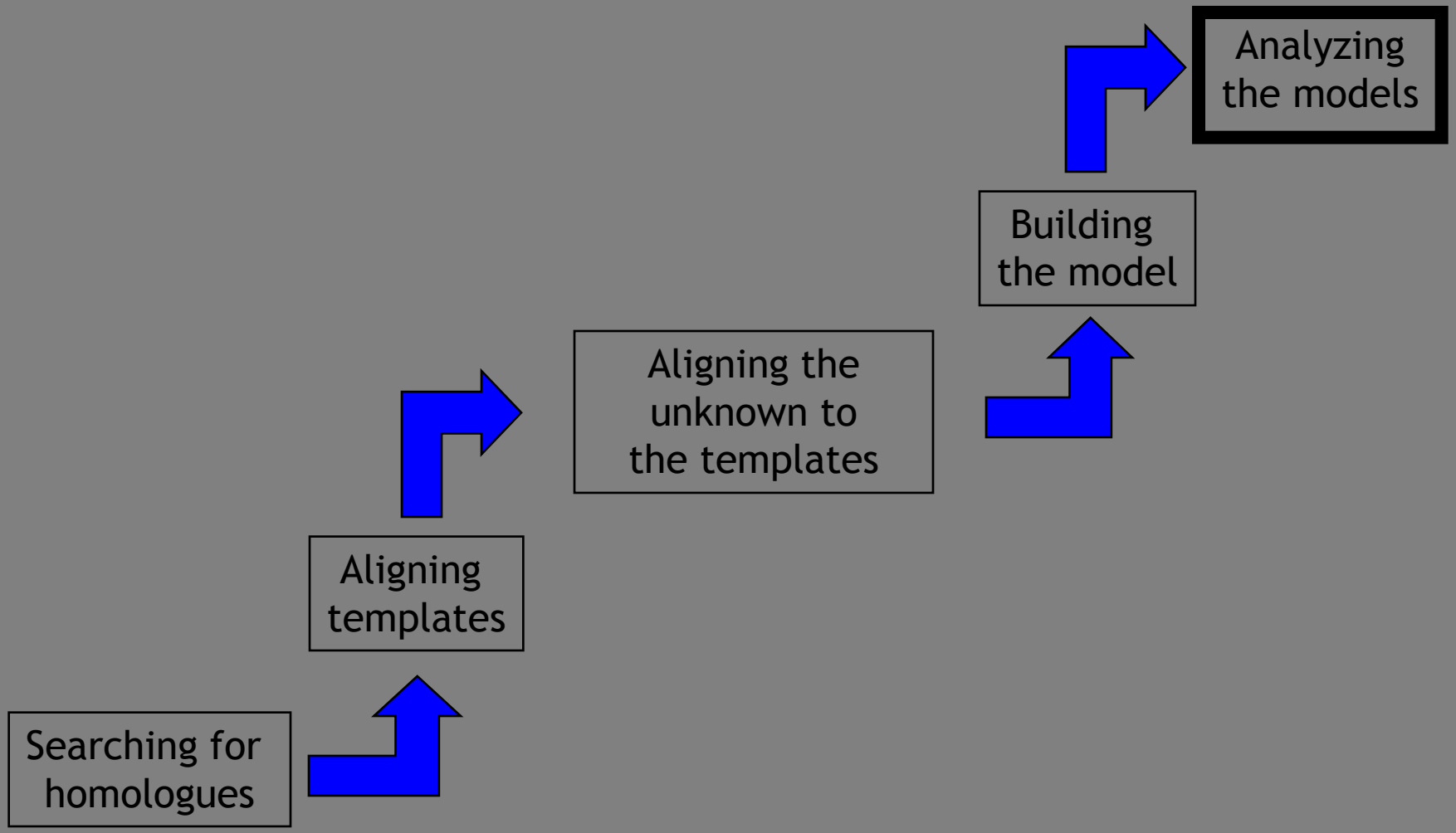
Hands-On Exercises

- Lesson 1: Modeling HDAC1 (Hom1)
  - Building of HDAC1 model
  - Load an alignment with associated structures
  - Use the Build Models protocol to run MODELER
  - Examine the Jobs Explorer
  - Load and display the results

Validation and Analysis

Is the model useful and what can it tell us?

Structure Prediction by Homology Modeling



Is the model correct?

- *Can* it be correct based upon the templates you have?
- Does this model “look” like a protein?
- Can it be used to address experimental data?
- How much confidence can we place in the model?
- How can I refine the model?

Can it be correct based upon the templates?

- The model can be no more accurate than the reference structures

- Validate your reference structures
  - What is the level of sequence identity to the unknown?
    - The higher the sequence identity, the better your model
  - Be aware of their strengths and weaknesses in the structure
    - Resolution of X-ray crystal structures
    - B-factors for important residues
  - Where are missing residues?
  - Analyze template structures and compare to your model
    - Profiles-3D
    - Ramachandran plot

What can MODELER provide?

- Model was created by optimizing the objective function
  - Trying to fit restraints in the model with the corresponding restraints in the templates
- Violations of restraints are kept
  - Differences between the target and what is in the model
  - Segregated per residue
- Violations summed for each residue
  - Gives an indication of the quality of the model in and around that residue
- The violations are useful in making comparisons between different models
- Referred to as a PDF energy

MODELER Analysis

- Measures are reported in Summary section of Report.htm

Summary

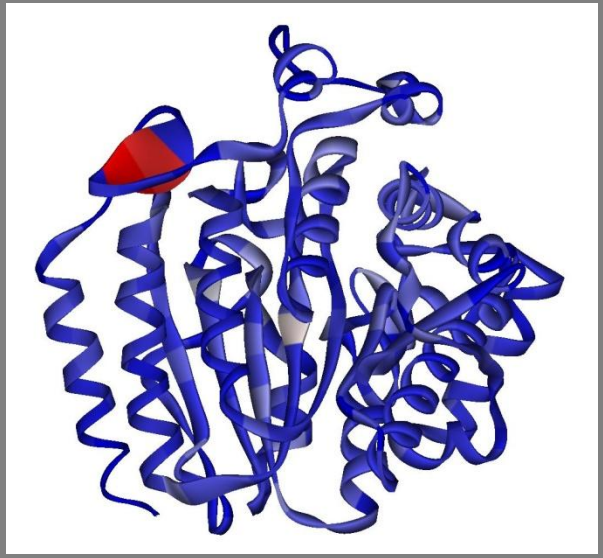
Models Sorted by PDF Total Energy

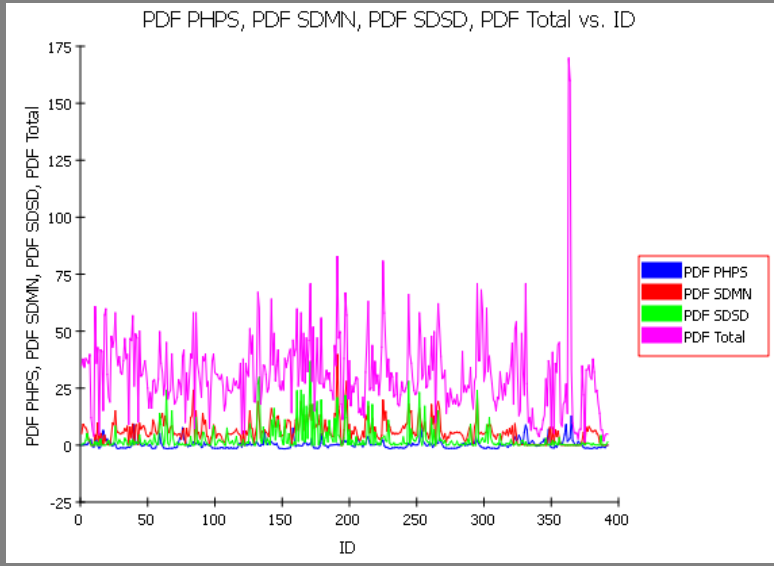
Model Name	PDF Total Energy	PDF Physical Energy	DOPE Score
HDAC1.B99990001	12015.24	1196.1	-44900.43
HDAC1.B99990002	12039.49	1199.74	-44706.15
HDAC1.B99990003	12113.72	1212.44	-44887.91

  - PDF Total Energy
    - Sum of all homology-derived and stereochemical pseudo-energy terms
  - PDF Physical Energy
    - Sum of energies of the stereochemical pseudo-energy terms and knowledge based non-bonded potentials
  - DOPE Score
    - An atomic based statistical potential measures the relative stability of a conformation

MODELER Analysis

- PDF terms per residue are loaded the model MSV file
  - Can be viewed through Data Table
  - 40 separate terms
    - Listed in the documentation for Analyzing results - Build Homology Models
- Used to evaluate models through the protein display
  - Color ribbon
  - Adjust size of ribbon
- Can also be graphed with the **Chart** menu





DOPE

- Available in **Verify Protein (MODELER)** protocol
  - Also found in other protocols
- Discrete Optimized Protein Energy
  - Statistical potential for model evaluation
  - Atomic based
  - Distance dependent
  - Derived from a set of known protein structures
- Can be considered as a conformational energy
  - Measures the relative stability of a conformation with respect to other conformations of the same protein
  - Assists in choosing the best model out of a set
  - Used as part of the energy function in predicting loop models or to optimize the local structure of a mutated residue

DOPE

- Algorithm
  - Considers pairwise probablity density functions (PDFs) for all atom pair types in a protein
    - Probabilities of two given atom types being within a certain distance of each other
  - Uses that information to calculate the joint PDF
    - Probability of a given atom being positioned in a native structure given the set of pairwise PDFs
  - Joint PDF can be related to the total free energy of the system
  - Interpolated using cubic splines for smooth function values
    - Renders DOPE useful for optimization

DOPE

- Can be used to detect the best fold from a set of protein conformations
  - Low score is best
- Caveats
  - Only can compare different folds of the same sequence
  - More accurate with near-native structures
  - Only operates on single chains
  - Best operates on monomers
    - May give misleading results with units from a multimer
  - Less accurate with:
    - Small proteins (<50 residues)
    - NMR structures

Is the model correct?

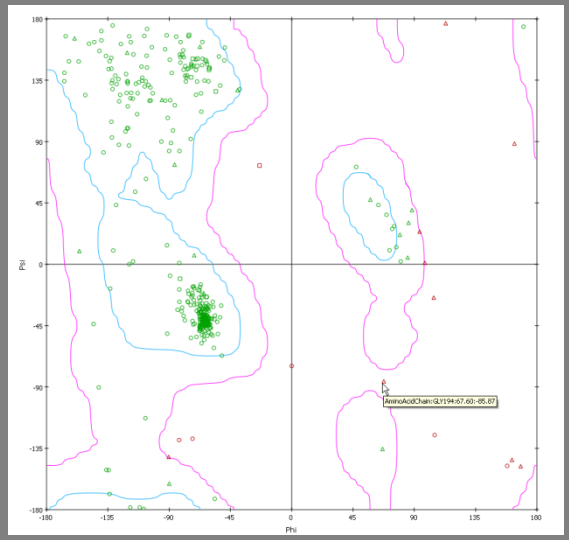
- *Can* it be correct based upon the templates you have?
- Does this model “look” like a protein?
- Can it be used to address experimental data?
- How much confidence can we place in the model?
- How can I refine the model?

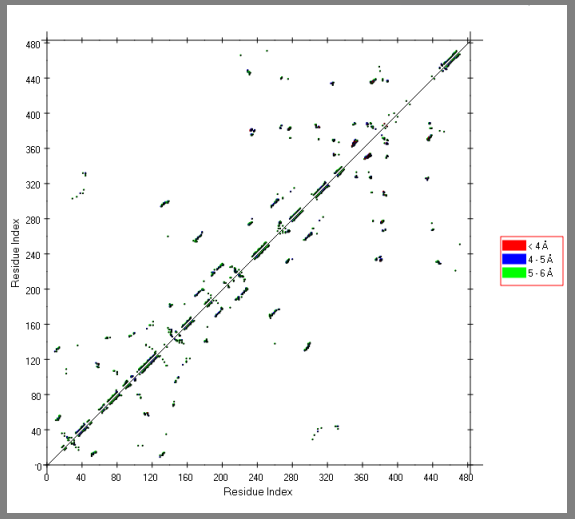
Does this model “look” like a protein?

- Examine properties of the protein

- From the **Chart** menu
  - **Ramachandran Plot**
    - Are any phi/psi angles in disallowed regions?
  - **Contact Plot**
    - Inter-residue contacts

- Exposed residues
  - Are there large hydrophobic patches on the surface?





Profiles-3D Analysis

- Given this fold or structure, is it compatible with the given sequence?
  - You have both the sequence and a proposed structure.
  - Now does it make sense?

- Run via the **Verify Protein (Profiles-3D)** protocol

- Scoring function based on residue environment
  - Calculate the compatibility of each residue in a sequence with its predicted 3D environment
  - Reduces the 3D structure to a 1D string of residue environments
- Examine scores on a per residue basis

Profiles-3D Residue Environment

- Six basic environments are categorized according to
  - Buried side chain area
  - Side chain area that is exposed to polar atoms
- Further subclassifications made based upon secondary structure
  - Yields 18 total environments
  - 18 environments for each of the 20 amino acids
  - Will consider residue position relative to implicit membrane when assigning environment class
- Residue position in a 3D structure is evaluated using a 3D-1D scoring function
  - Relating the compatibility of a given residue in a given environment
- Sum of the 3D-1D residue score gives the Verify Score for the predicted protein structure
  - The Verify Score is then normalized by the length of the sequence

Profiles-3D Analysis

- Statistical analysis in Report.htm and Output.log file

Summary

HDAC1 Verify Score = 151.46

HDAC1 Verify Expected High Score = 178.758

HDAC1 Verify Expected Low Score = 80.441

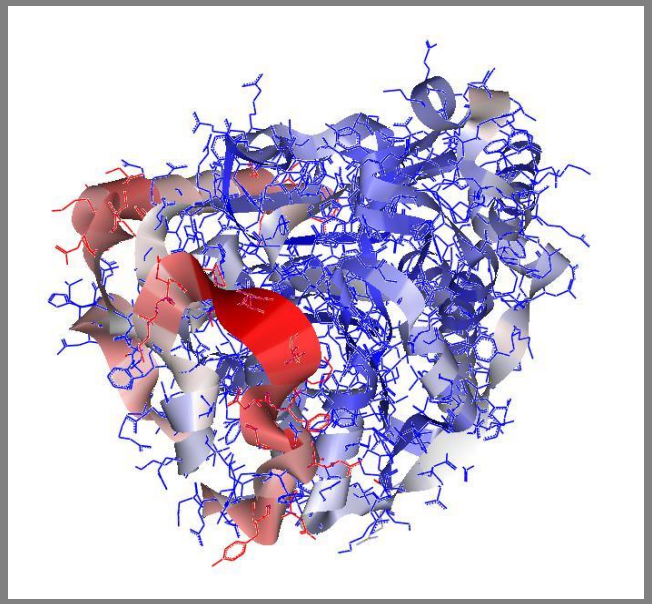
- Data contained in DSV file in Output folder

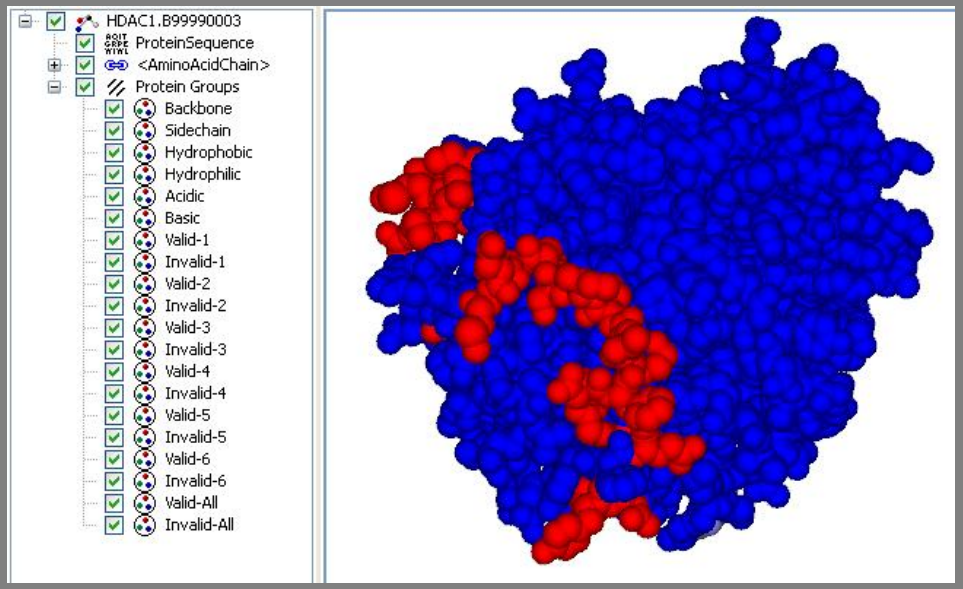
	PDF SSVA	PDF STOR	PDF Total	Verify Buried Area	Verify Environment Class	Verify Polar Fraction	Verify Score
1	0	0.82	28	7.588	E	0.937	0.753
2	0	1.5	31	96.295	P2	0.75	0.753
3	0	1.7	31	125.189	B1	0.266	0.753
4	0	1.9	33	56	P1	0.37	0.753
5	0	3.4	32	193.857	B1	0.198	0.753
6	0	3	33	195.931	B2	0.383	0.753
7	0	3.1	47	137.025	B3	0.544	0.857

Molecule | ProteinSequence | AminoAcidChain | AminoAcid | Atom | Bond | Group

Profiles-3D Analysis

- Data displayed as ribbon
  - Coloring by residue score
- Groups specified for Valid and Invalid residues





Is the model correct?

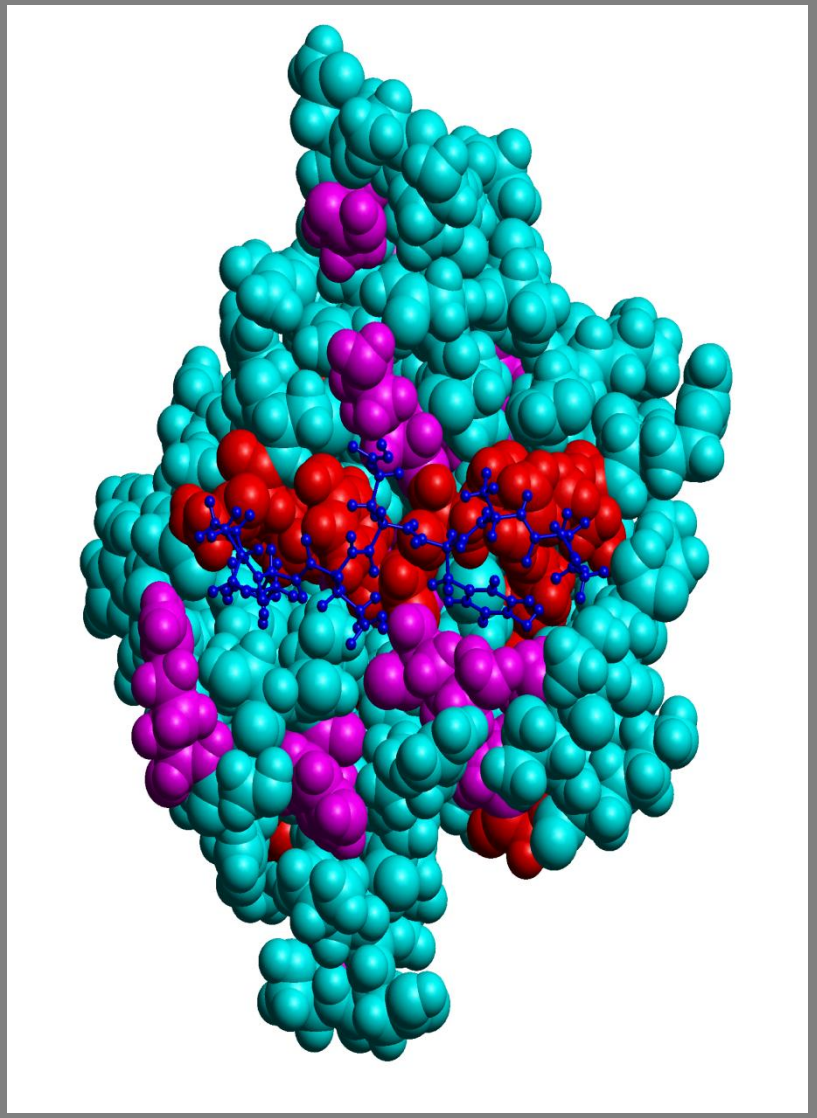
- *Can* it be correct based upon the templates you have?
- Does this model “look” like a protein?
- Can it be used to address experimental data?
- How much confidence can we place in the model?
- How can I refine the model?

Use It!

- Put the model to work and see how it succeeds
  - See if it agrees with known information
- Are known functionally important residues in the correct positions?
  - Catalytic site
  - Binding site
  - Protein-protein interactions
- Does the model explain the affects of specific mutations?
  - Examine positions of mutants in model with Evolutionary Trace

Evolutionary Trace

- Method implemented based on the work of Prof. Olivier Lichtarge at Baylor College of Medicine, Houston
- Analyzes a multiple sequence alignment to identify sequence families
- Derives trace residues based on consensus sequences from clustered families
  - Conserved
    - Conserved in all sequences
    - Affect function
  - Class-specific
    - Conserved within clusters
    - Affect specificity and activity

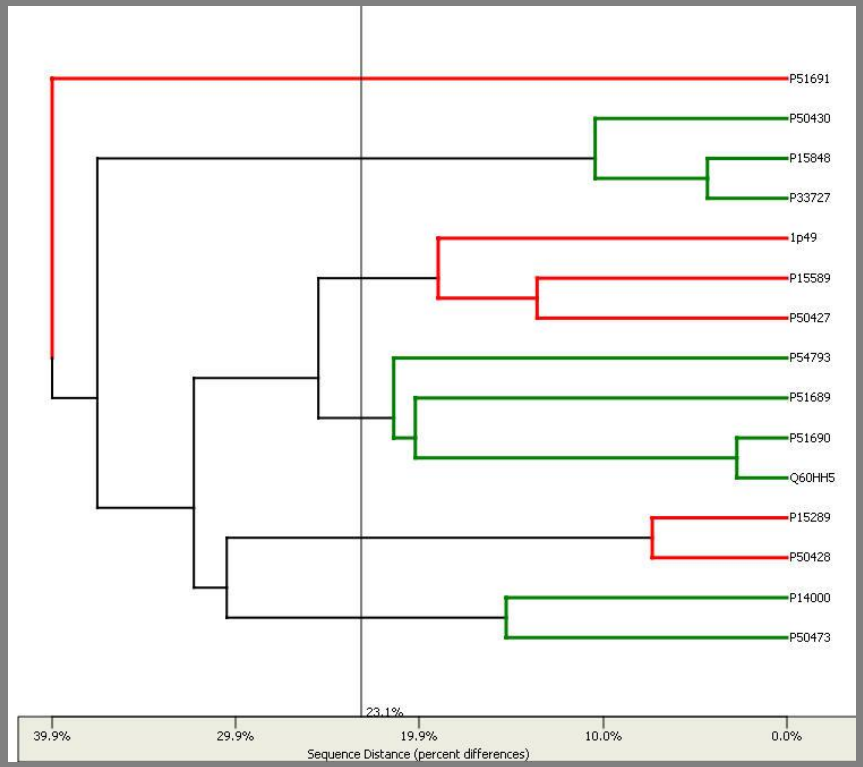


Evolutionary Trace

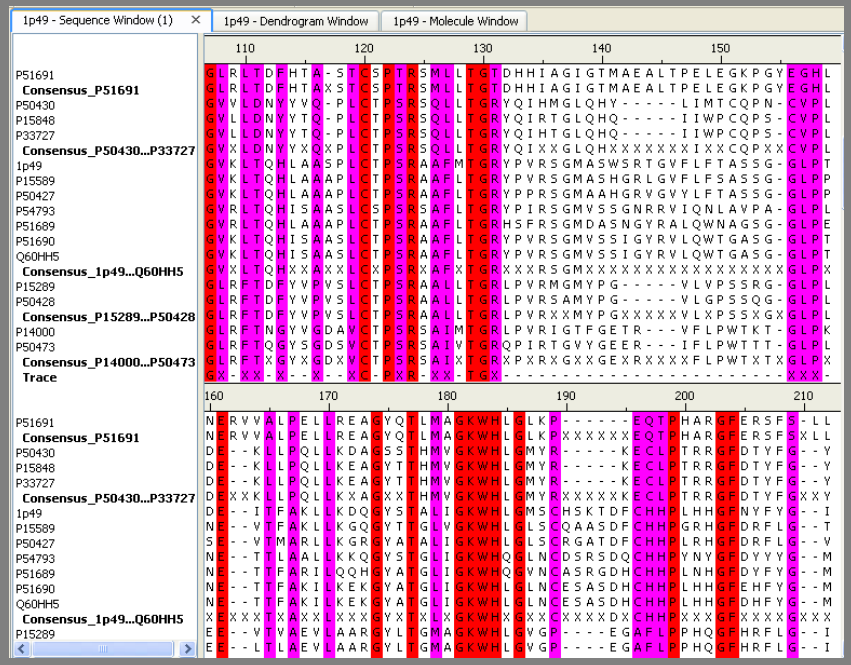
- Requires a set of related sequences as input
  - Minimum of 15 is recommended
  - Sequence identity in the range of 20-80%
- Alignment produced by any reliable method
  - Can be read in via an alignment file
  - Can be created with an alignment protocol
    - For example: PSI-BLAST output or the Multiple Sequence Alignment protocol
- Evolutionary Trace tools
  - Identifies trace residues
  - Performs clustering

Evolutionary Trace Dendrogram

Adjust vertical line to adjust clustering distance



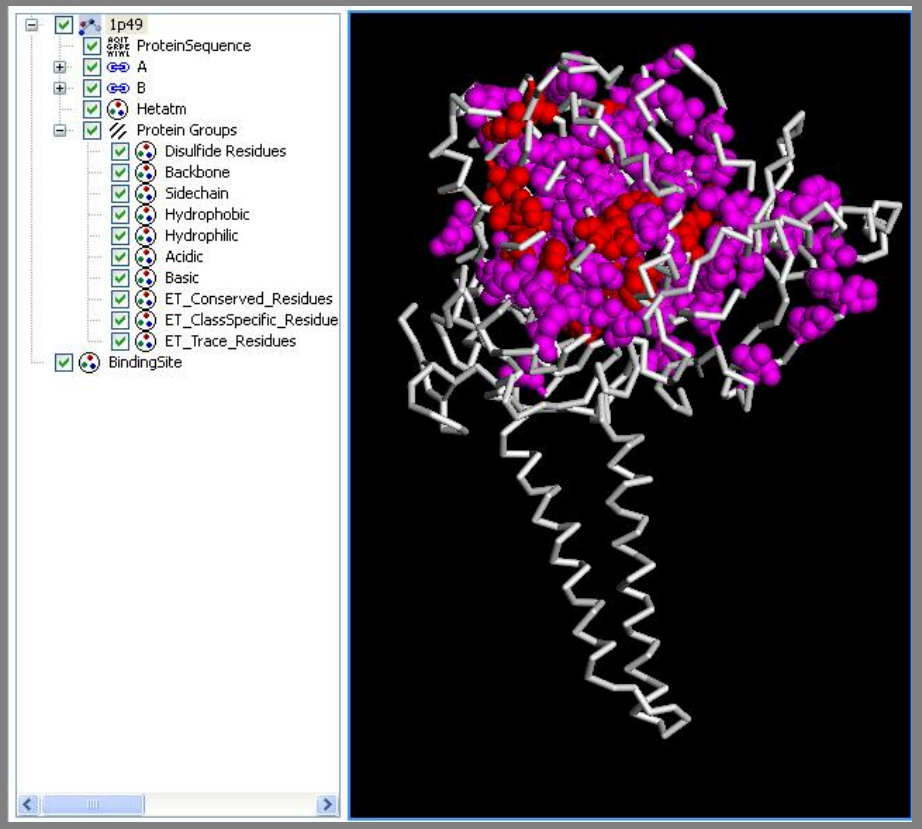
Window is updated as clustering changes



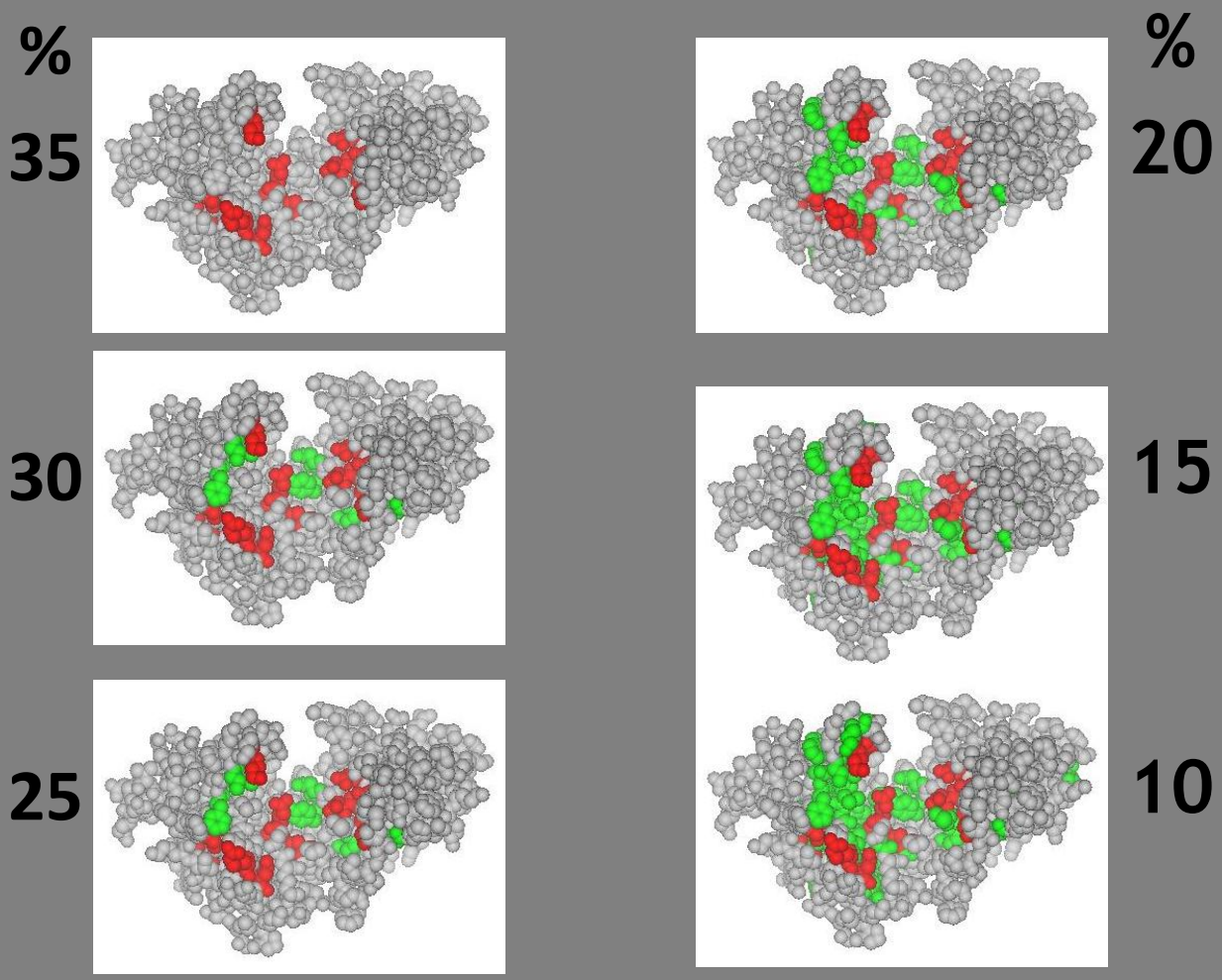
Coloring changed via Display Style command

Visualization of Trace Residues

- If a structure is available, trace residues may be mapped in 3D
- Groups are defined for conserved and class-specific residues
  - Groups updated as sequence dendrogram is updated
- Can be clustered to identify significant patches
  - **Analyze Structures – Evolutionary Trace** tool group



HIV-1 Reverse Transcriptase

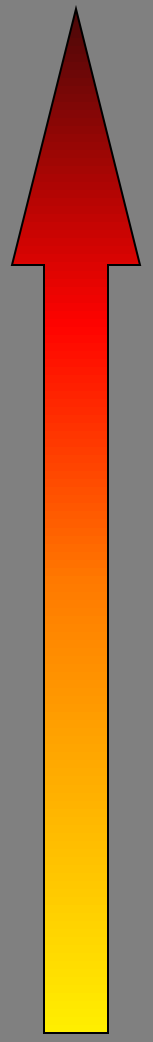


Is the model correct?

- *Can* it be correct based upon the templates you have?
- Does this model “look” like a protein?
- Can it be used to address experimental data?
- How much confidence can we place in the model?
- How can I refine the model?

Sources of Errors in Homology Models

- Incorrect templates
  - Problem worsens with decreasing sequence identity
  - Use profile-based searches
- Incorrect alignments
  - Largest source of error at low sequence homology
  - Use multiple sequence and profile-profile alignments
- Regions without a template
  - Most difficult region to model
  - Some techniques are useful for small inserts
- Backbone distortions in correctly aligned regions
  - Different backbone traces due to sequence differences
- Side-chain packing
  - Can be important in active sites or binding sites



How can we minimize errors in our model?

- Look for errors in the reference structures
- Do the best job we can on the sequence alignment
  - Alignment is the major source of error at less than 30% sequence identity
- Validate the final model thoroughly
- Try to correlate observed experimental data with the model
- Always treat the homology model as a working hypothesis

Refining the Model

"Refinement remains the principal bottleneck to progress."
John Moult, 2005

- Becomes more of a problem with:
  - Decreasing sequence homology between model and template
  - Similarity to native structures
- Energy minimization
  - Overall, brute-force minimization may not be the best approach
  - Feig and Brooks (2002)
    - Short minimization with Cα restraints followed by short unrestrained minimization improves results
    - Better performance with distance-dependent dielectric constant and ε = 2 or 4

Refining the Model

- Molecular dynamics
  - Flohil et al. (2002)
    - Constrained well refined portions of the model
    - Ran nanosecond MD with explicit solvent
    - 15–20% reduction in RMSD imperfections of model relative to known structure
  - Chen and Brooks (2007)
    - Include reliable structural information
      - Restraints for secondary structure and internal contacts
    - Use 3-5 nanosecond molecular dynamics
      - Sampling with replica exchange
      - Use generalized Born implicit solvent model
    - Indications are that good refinement can occur when the initial structure is a near-native structure (3-5 Å from the actual structure)

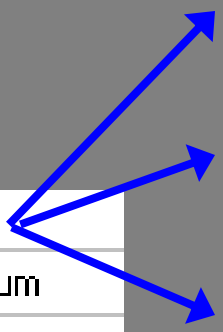
Local Refinements of the Model

- Loop modeling and refinement
  - While building the models with the **Build Homology Models** protocol
    - **Refine Loops** parameter creates multiple models
    - No control over which loops to refine
  - Post processing of models
    - **Loop Refinement (MODELER)** protocol
      - Same algorithm as in **Build Homology Models** protocol
      - User specifies what loops to refine
    - **Loop Refinement** protocol
      - CHARMm-based method named LOOPER
- Side-chain refinement
  - ChiRotor
    - CHARMm-based method

Loop Modeling and Refinement

- In the **Build Homology Models** protocol...
- Creates additional models as loops are refined
  - **Number of Models** parameter specifies the number of basic models created
  - **Refine Loops | Number of Models** creates additional models from each basic model
- For example, with HDAC1 models

⊞ Number of Models	3
Optimization Level	Medium
⊟ Refine Loops	True
Number of Models	2
Optimization Level	Medium



HDAC1.B99990001.dsv
HDAC1.BL00010001.dsv
HDAC1.BL00020001.dsv
HDAC1.B99990002.dsv
HDAC1.BL00010002.dsv
HDAC1.BL00020002.dsv
HDAC1.B99990003.dsv
HDAC1.BL00010003.dsv
HDAC1.BL00020003.dsv

Local Refinements of the Model

- Loop modeling and refinement
  - While building the models with the **Build Homology Models** protocol
    - **Refine Loops** parameter creates multiple models
    - No control over which loops to refine
  - Post processing of models
    - **Loop Refinement (MODELER)** protocol
      - Same algorithm as in **Build Homology Models** protocol
      - User specifies what loops to refine
    - **Loop Refinement** protocol
      - CHARMm-based method named LOOPER
- Side-chain refinement
  - ChiRotor
    - CHARMm-based method

Loop Refinement (MODELER) protocol

- Post-processing of homology models
- Must specify a range of residues encompassing a loop
  - Selected in Graphics View
  - Can have more than one loop region
- Uses loop generation algorithm to create multiple loops

Local Refinements of the Model

- Loop modeling and refinement
  - While building the models with the **Build Homology Models** protocol
    - **Refine Loops** parameter creates multiple models
    - No control over which loops to refine
  - Post processing of models
    - **Loop Refinement (MODELER)** protocol
      - Same algorithm as in **Build Homology Models** protocol
      - User specifies what loops to refine
    - **Loop Refinement** protocol
      - CHARMm-based method named LOOPER
- Side-chain refinement
  - ChiRotor
    - CHARMm-based method

Loop Refinement Protocol with LOOPER

- Operates only on the defined loop region
  - Remainder of protein is fixed
- Three stage algorithm
  - Loop divided into two halves
    - Systematic search of ϕψ angles in loop region
      - Uses preferred combinations
    - Each loop half ranked by CHARMm energy
    - More favorable halves are linked and minimized
  - Side chains are placed using ChiRotor
    - Minimized
    - Ranked by CHARMm energy
  - Solvation term added
    - Loops rescored by CHARMm energy with a generalized Born solvation model

LOOPER Performance

- Test set of protein structures as defined by Fiser *et al.*
  - Loops ranging in length from 2-12 residues
  - 40 loops of each length from different proteins
- RMSD reasonable up to 9-10 residues
  - Compute time in minutes
- For 8-residue loops...
  - 75% less than 2 Å RMSD relative to crystal structure

“Knots” from MODELER

- Common complaint about MODELER
- Occur due to insufficient information for placement of chain
  - Usually when one or more neighboring long insertions (longer than 15 residues) have no match in template
- Look for problem areas in templates
  - Undefined residues
  - Missing loops
  - Terminal overhang

Repairing “Knots”

- Reconsider the alignment
- Cut overhangs
  - On by default in the **Build Homology Models** protocol
- Calculate independently many loop models
  - Set the **Refine Loops** parameter to *True*
  - Specify **Number of Models** parameter
  - Simply eliminate the knots in selecting the best representatives
- Use the loop refinement protocols
  - Post-processing of built models
- Model longer loops as independent folding domains
  - Use distance restraints to keep the termini at a reasonable orientation relative to tail region

Local Refinements of the Model

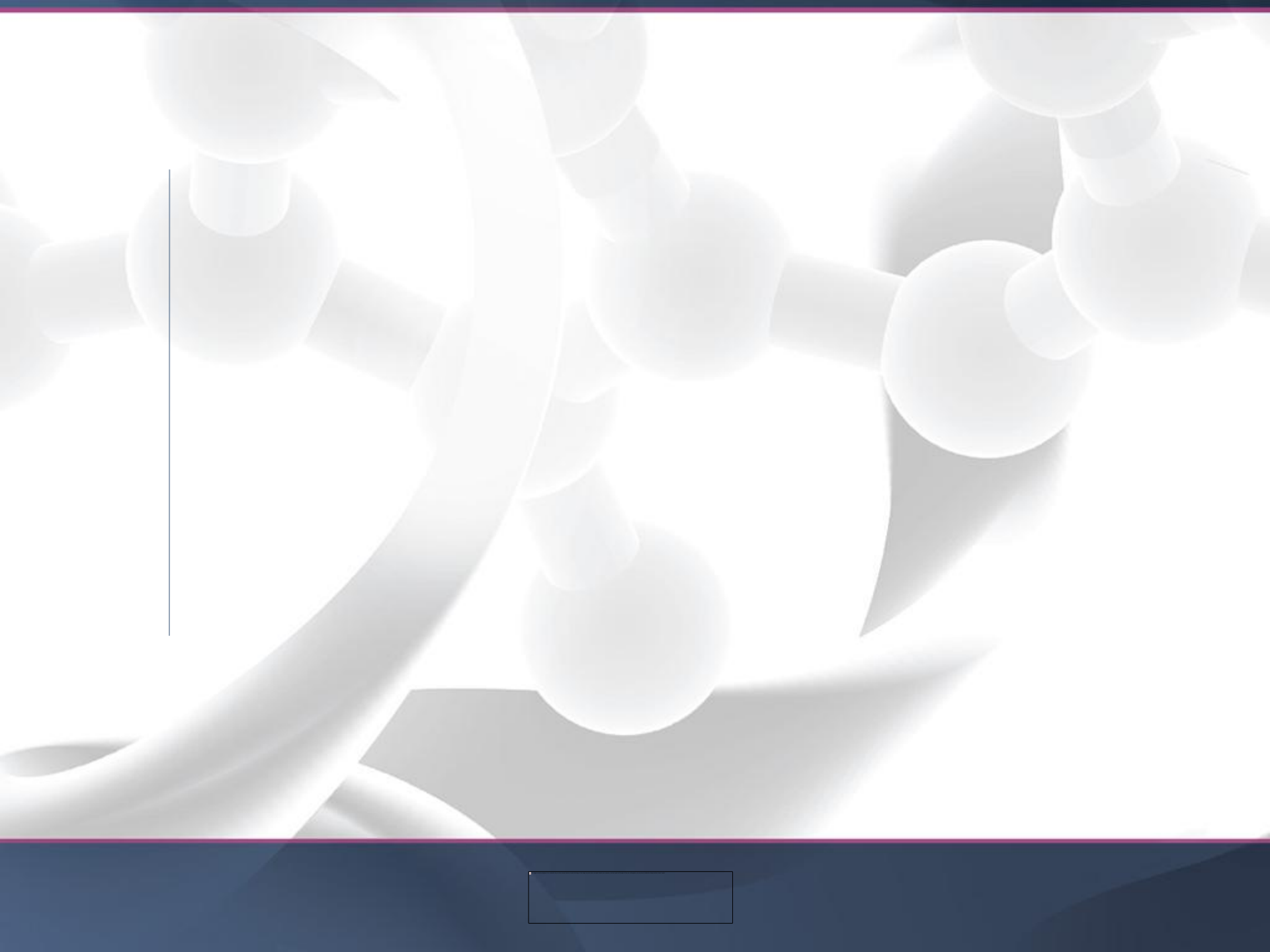
- Loop modeling and refinement
  - While building the models with the **Build Homology Models** protocol
    - **Refine Loops** parameter creates multiple models
    - No control over which loops to refine
  - Post processing of models
    - **Loop Refinement (MODELER)** protocol
      - Same algorithm as in **Build Homology Models** protocol
      - User specifies what loops to refine
    - **Loop Refinement** protocol
      - CHARMm-based method named LOOPER
- Side-chain refinement
  - ChiRotor
    - CHARMm-based method

Side-Chain Refinement Protocol

- Optimizes the protein side-chain conformation
- Based on ChiRotor algorithm
  - Systematic searching of side-chain conformation
  - CHARMm energy minimization
  - Requires:
    - Properly force field typed protein using one of the CHARMm force fields
    - Selected residues to optimize
- Returns one protein structure with optimized side chains
- Ignores alanine, glycine, and disulfide bonds

Side-Chain Refinement Protocol

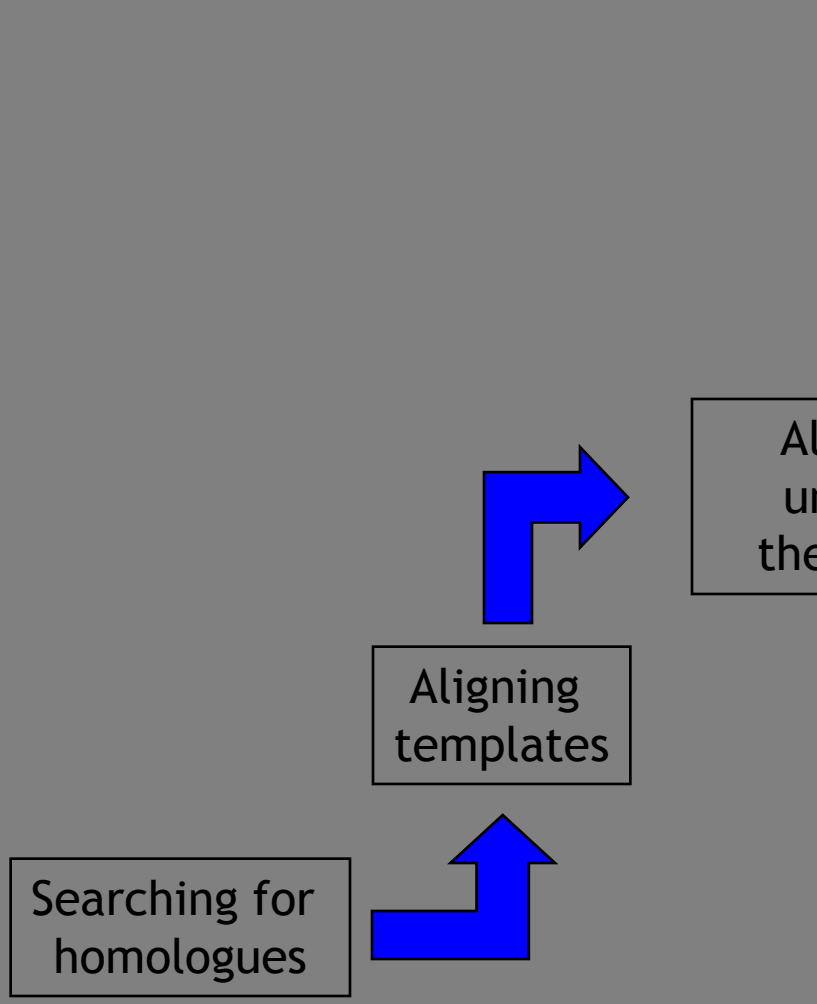
- Algorithm
  - Remove side chains from selected residues
  - Each residue χ_1 torsion sampled in three states
    - Optimized with CHARMm and two lowest conformations retained
    - All other side chains to sample are removed
    - All other protein atoms are fixed
  - Lowest energy side chain selected for each residue
    - Minimized
    - Tested with second conformation one residue at a time
      - Minimized
      - If lower in energy side chain kept



Is the model correct?

- Use MODELER PDF energies and DOPE scores
- Examine the model structure and compare to templates
  - Ramachandran plot
  - Profiles-3D score
- Test the model with experimental data
- Refine the model if needed
  - Redo the sequence alignment and rebuild model
  - Energy minimization and/or molecular dynamics
  - Loop refinement
  - Side-chain refinement

Structure Prediction by Homology Modeling



- Thoroughly study the model
- Use MODELER PDF energies and DOPE score
- Compare models to template
- Can it address experimental data?
- Refine model as needed

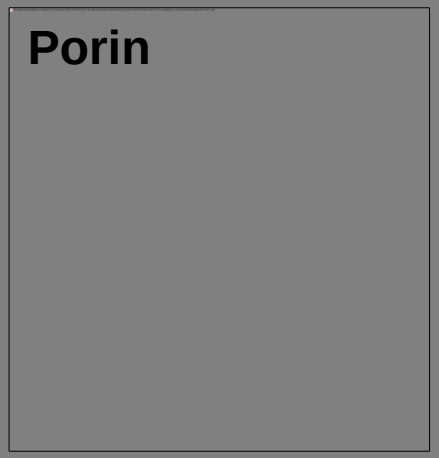
Hands-On Exercises

- Analysis and characterization
- Lesson 1: Examining the HDAC1 Models (Valid1)
  - Load the completed homology model of HDAC1
  - Examine the data provided by MODELER
  - Display the data as a ribbon trace
- Lesson 2: Profiles-3D and the Verify Score (Valid2)
  - Analyze HDAC1 model with Profiles-3D
- Lesson 3: Loop Refinement (Valid3)
  - Using LOOPER algorithm on the HDAC1 model

Transmembrane Modeling

Transmembrane Proteins - Background

- Several types of proteins have membrane domains



GPCR

- Most popular drug targets
- Responsible for signal transduction
- 7 helical transmembrane domain
- Many have ligand binding site in transmembrane domain
- Hard to crystallize, only 3 known structures
- Diverse sequences

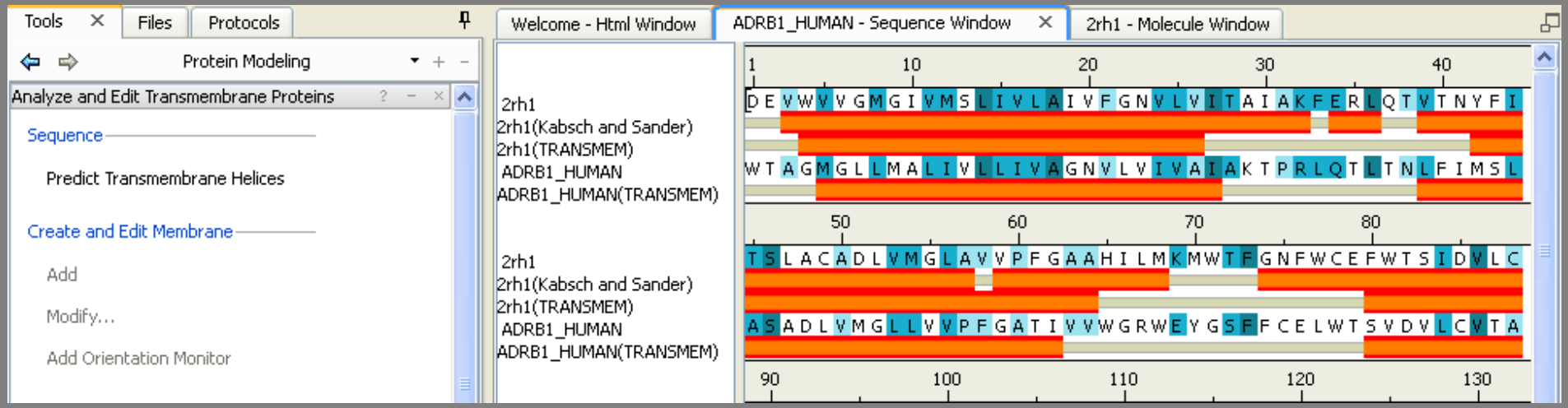
Transmembrane Modeling in DS

DS Tools and Protocols can be used to:

- Predict helical transmembrane domains from input sequence
- Align sequences using transmembrane predicted secondary structure
- Add membrane object to transmembrane domain of protein
- Visualize the membrane object along with the protein structure
- Manually adjust the membrane object orientation
- Optimize the membrane object orientation
- Run Simulations using implicit membrane solvation model

Predicting Transmembrane Helices from Protein Sequence

- **Predict Transmembrane Helices** found in **Analyze and Edit Transmembrane Proteins** tool panel
- Uses TMHMM (*Transmem* from **Wisconsin package**)
- **Sequence | Secondary Structure | Edit** allows you to remove the secondary structure of the globular part from PDB or Kabsch-Sander



Use Predicted Transmembrane Helices to assist in Sequence Alignment

- Use **Align Multiple Sequences** protocol from the **Sequence Analysis** folder
- **Use Secondary Structures** parameter and choose *TRANSMEM*

Add an Implicit Membrane to a Structure

Membrane: represents a planar low-dielectric slab to approximate the non-polar region of a lipid bilayer

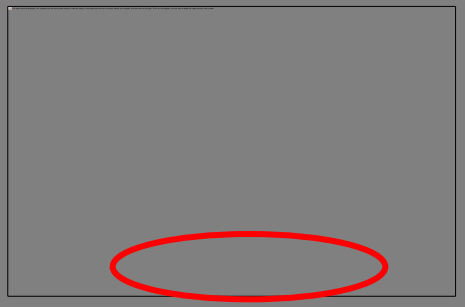
- **Quick Method**
  - Use **Add** command from **Analyze and Edit Transmembrane Proteins** tool panel
- **Rigorous Method**
  - **Add Membrane and Orient Molecule** protocol

Implicit Membrane – Using 'Add' Tool

- Default thickness 30 Angstrom
  - Modify using **Transmembrane Protein** subpage of Preferences
    - Only if experimental data indicates a different thickness
  - 30 Angstrom thickness represents general biological membranes
- Two parallel planes represent the membrane in the Graphics View
  - Hydrophobic region of the membrane corresponds to space between the planes
  - Blue plane represents side of membrane with more positively charged residues. In most proteins it corresponds to the cytoplasmic side (positive inside rule)
  - Other plane is green
  - If no significant difference in charged residues both planes are green

Implicit Membrane – Using ‘Add’ Tool

- A membrane object appears in the Hierarchy View



- Membrane placement is determined by optimizing a simplified solvation energy
- Modify tool allows for manual adjustments to thickness and position of membrane

Using *Add Membrane and Orient Molecule* protocol

- Uses a stepwise search algorithm for optimizing the orientation of the membrane
- Optimal orientation corresponds to the minimum solvation energy and relates to three variables:
  - Angle of molecule rotation around one of the principle axis of inertia
  - Tilt angle of the selected axis relative to the membrane normal
  - Distance from the molecular center of mass to the membrane mid-plane
- Creates the same type of Membrane object as the **Add Membrane** tool

Using *Add Membrane and Orient Molecule* protocol

- Solvation energy can be calculated using two similar approaches (for details see references)
  - GBIM: The non-polar region of the membrane is approximated as a planar dielectric slab having the same dielectric constant as inside the molecule.
  - GBSW: Consistent with continuum Poisson-Boltzmann (PB) electrostatics, the membrane is approximated as an solvent-inaccessible infinite planar low-dielectric slab
- Non-polar surface term
  - For non-transmembrane segments (i.e. peptides) surface term is essential
    - For transmembrane proteins may turned off
  - With GBIM - takes significant extra time, not worthwhile
  - With GBSW – not much extra CPU time, include in calculation

Getting a good placement…

FAST

- Use the tool first followed by the protocol
- The ‘Add’ tool can make a good ‘first guess’ of placement
- Use the ***Add Membrane and Orient Molecule*** protocol with a membrane created using the ‘Add’ tool
  - Reduce the Shift From = -5 and To = 5
  - Reduce the Tilt Angle Maximum Change = 30, Step Size = 5
  - Will greatly reduce computation time of the protocol

Getting a good placement…

ACCURATE

- First perform the steps in the previous ‘FAST’ method
- Run ***Calculate Protein Ionization and Residue pK*** to reprotonate the protein
- Run ***Add Membrane and Orient Molecule*** on the results from ***Calculate Protein Ionization and Residue pK*** to refine membrane orientation on reprotonated protein

Protocols that use Implicit Membrane

- Protocols will use the membrane information automatically if there is a membrane added to the input molecule
- Verify Protein (Profiles-3D)
- Protein Ionization and Residue pK
  - GBIM and CHARMm polar and CHARMm
- Loop Refinement - GBIM
- Most of the simulation protocols, e.g. Minimization, Dynamics, etc

Protocol Parameters and Settings

Protocols	Solvation Model	ForceField	Suggested Membrane thickness	Dielectric Constant	Nonpolar SA term
Calculate Protein Ionization and Residue pK	GBIM	CHARMm & CHARMm polar	30	10	no
Loop Refinement	GBIM	CHARMm & CHARMm polar	30	2	no
Add membrane and orient molecule	GBIM and GBSW	All CHARMm forcefields	30	2	yes
Other simulations	GBIM and GBSW	All CHARMm forcefields	30	2	yes

Verify Protein (Profiles-3D) - Membrane

- Profiles-3D –
  - 18 environment classes based on:
    - Sidechain buried (exposed) area
    - Fraction polar (FP) – fraction of the sidechain in contact with polar environment
      - Polar atom in protein
      - Solvent
    - Secondary structure
- Profiles-3D Membrane – adjust the following for assigning environment classes
  - Sidechain exposed area = 0 for residue in membrane
  - Fraction polar (FP’ = FP – sidechain exposed area)
  - Best to use 30 A thickness for membrane

Protein Ionization and Residue pK - Membrane

- Accounts for the effect of a lipid membrane on the ionization characteristics of the titratable groups
- Membrane is modeled as a planar low-dielectric slab using the Generalized Born solvation model with implicit membrane (GBIM)
- No specific parameter settings when running protocol, automatically deployed when membrane object is present

pK Prediction - Bacteriorhodopsin 1c3w.pdb[1]

	$pK_{1/2}$			
	Calculated GBIM	**Calculated without membrane**	**Calculated using MEAD with PB and membrane[2]**	**Experiment**
ARG82	**> 14**	**>14**	**>15**	**>13.8**
ASP85	**2.96**	**7.1**	**1.7**	**2.6**
ASP96	**8.80**	**8.7**	**>15**	**>12**
ASP115	**6.54**	**8.1**	**8.4**	**>9.5**
GLU194	**9.69**	**8.6**	**> 15**	**Proton release dyad; pK~9.5**
GLU204	**3.35**	**8.7**	**< 0**	
ASP212	**<0.00**	**7.1**	**<0**	**<2.5**
Schiff base216	**> 14**	**12.1**	**>15**	**>12**

[1]Luecke et al. (1999) *J. Mol.Biol.*,291,899-911.
[2]Spassov et al. (2001) *J. Mol.Biol.*,312,203-219

pK Prediction - Beta2-adrenergic receptor

Calculated pK values of the residues inside the membrane

	$pK_{1/2}$	
	40 Angstr. membrane	Without membrane
Asp79	7.37	9.47
ASP113	8.07	9.52
GLU107	9.44	6.66
GLU122	10.69	7.30
HIS296	5.43	6.7 3

Cherezov et al. (2007) *Science* ***318:*** *1258-1265*

Simulations

- Can be used with any Simulation protocol containing an **Implicit Solvent Model** parameter in which GBSW and GBIM are valid choices
- Input molecule must contain a Membrane object
- When possible use the same Solvation parameter values to run the simulation as those which were used to set up the system for simulation:
  - Add/orient the membrane (with ***Add Membrane and Orient Molecule***)
  - Protonate the protein (with ***Calculate Protein Ionization and Residue pK***)

Loop Refinement

- Include the effect of a lipid membrane on the loop conformation
- Requires the presence of a Membrane object in the *Input Typed Protein Molecule*
- Hardcoded to use GBIM solvation model with no Non-polar Surface Area term
- The protocol has no explicit parameters to set to specify the inclusion of a membrane

Transmembrane Protein - Resources

- GPCRDB
  - Alignment
  - Ligand binding data
  - Mutation data
  - http://www.gpcr.org/7tm/
- Pfam
  - Alignment
- SwissProt
  - TM helix definition
- Paper:
  - GBIM: Spassov VZ, Yan L, Szalma S; (2002) *J. Phys. Chem.* B106:8762-8738
  - GBSW: Im W.; Lee M.;, Brooks C.; (2003), J. Comput. Chem. 24(14):1691-1702
  - Profiles-3D: Protein Sci. (2001), 10(8): 1529–1538.