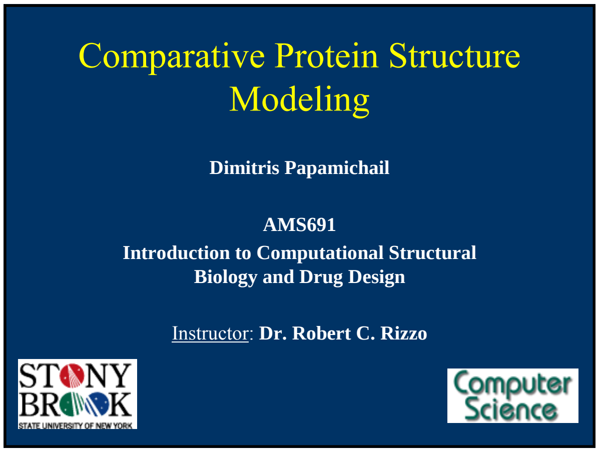
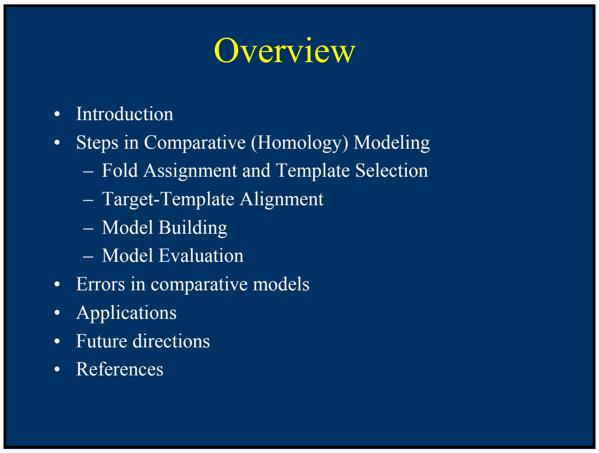


Comparative Protein Structure Modeling

Dimitris Papamichail

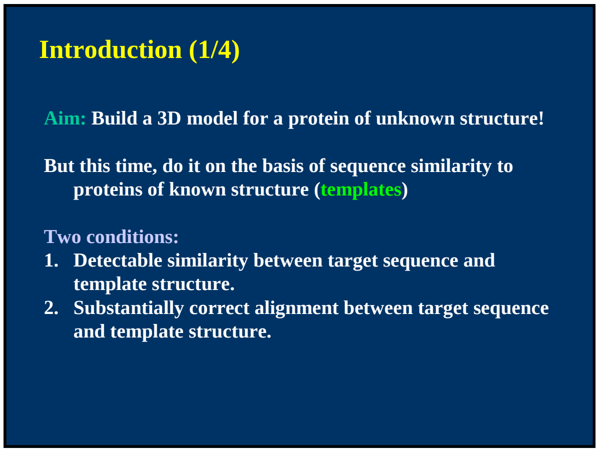
AMS691
Introduction to Computational Structural Biology and Drug Design

Instructor: **Dr. Robert C. Rizzo**



Overview

- Introduction
- Steps in Comparative (Homology) Modeling
  - Fold Assignment and Template Selection
  - Target-Template Alignment
  - Model Building
  - Model Evaluation
- Errors in comparative models
- Applications
- Future directions
- References



Introduction (1/4)

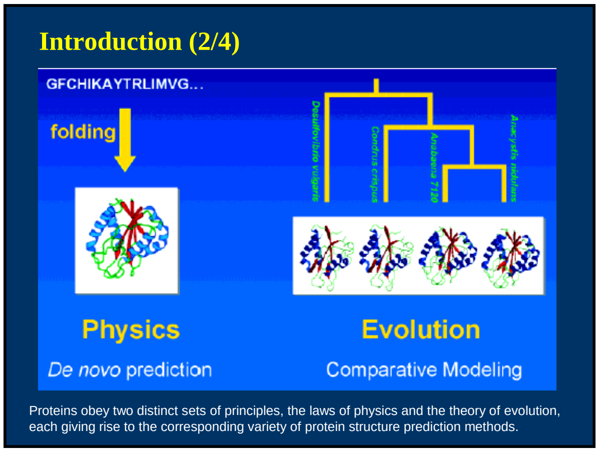
Aim: Build a 3D model for a protein of unknown structure!

But this time, do it on the basis of sequence similarity to proteins of known structure (templates)

Two conditions:

1. **Detectable similarity between target sequence and template structure.**
2. **Substantially correct alignment between target sequence and template structure.**

Introduction (2/4)



Proteins obey two distinct sets of principles, the laws of physics and the theory of evolution, each giving rise to the corresponding variety of protein structure prediction methods.

Introduction (3/4)

Why does it work?

Small changes in protein sequence usually result in small changes in the 3D structure

In other words:

The 3D structures of proteins from the same family are more conserved than their primary sequences.

Or again:

If similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed.

Even proteins with non-detectable sequence similarity can have similar structures.

Introduction (4/4)

Why use comparative modeling?

Most accurate prediction method! More accurate than ab initio protein structure prediction.

Overall accuracy of comparative models spans a wide range:

1. Low resolution models: Only correct fold.
2. Accurate models: Comparable with medium resolution structures determined by crystallography or NMR spectroscopy.

Even low resolution is useful, since aspects of functions can sometimes be predicted only from coarse structural features of a model.

Introduction (5/4)

Estimated that approximately 1/3 of all protein sequences are recognizably related to at least one known protein structure.

~ 600,000 known protein sequences =>
200,000 where comparative modeling can be applied.

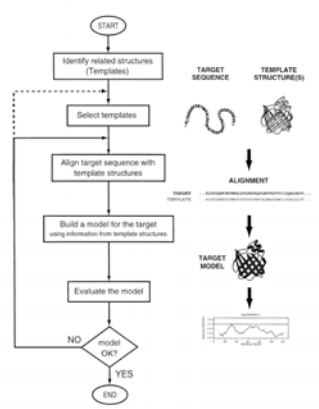
Usefulness of comparative modeling is increasing, because:

1. There is a limited number of unique structural folds.
2. Experimentally determined new structures increase exponentially.

Estimated that in less than 10 years at least one example of most structural folds will be known.

Steps in comparative modeling

1. Template Identification and Selection
2. Target-Template Alignment
3. Model Building
4. Model Evaluation


START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
model OK?
NO
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
ALIGNMENT
TARGET
TEMPLATE
TARGET MODEL

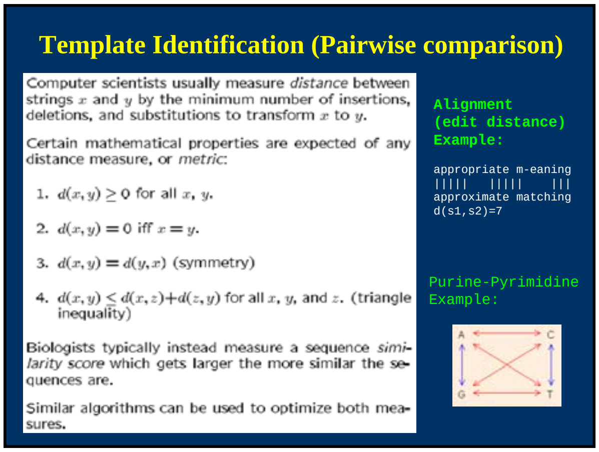
Template Identification and Selection

Where can we find templates?
Structural databases: PDB, SCOP, DALI, CATH

Depending on the genome, probability of finding related protein of known structure: 20% - 50%

Three comparison methods for fold identification:

1. Pairwise sequence-sequence comparison
2. Multiple sequence comparison
3. Threading



Template Identification (Pairwise comparison)

Computer scientists usually measure *distance* between strings x and y by the minimum number of insertions, deletions, and substitutions to transform x to y.

Certain mathematical properties are expected of any distance measure, or *metric*:

1. $d(x, y) \geq 0$ for all x, y.
2. $d(x, y) = 0$ iff $x = y$.
3. $d(x, y) = d(y, x)$ (symmetry)
4. $d(x, y) \leq d(x, z) + d(z, y)$ for all x, y, and z. (triangle inequality)

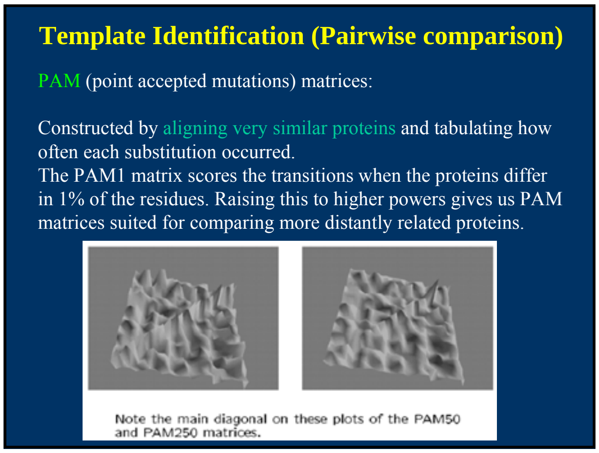
Biologists typically instead measure a sequence *similarity score* which gets larger the more similar the sequences are.

Similar algorithms can be used to optimize both measures.

Alignment (edit distance) Example:

```
appropriate m-eaning
|||||   |||||    |||
approximate matching
d(s1,s2)=7
```

Purine-Pyrimidine Example:

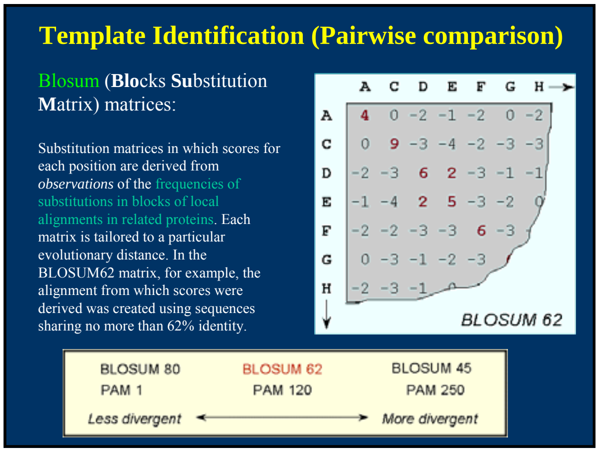


Template Identification (Pairwise comparison)

PAM (point accepted mutations) matrices:

Constructed by aligning very similar proteins and tabulating how often each substitution occurred.
The PAM1 matrix scores the transitions when the proteins differ in 1% of the residues. Raising this to higher powers gives us PAM matrices suited for comparing more distantly related proteins.

Note the main diagonal on these plots of the PAM50 and PAM250 matrices.



Template Identification (Pairwise comparison)

Blosum (**Blo**cks **Su**bstitution **M**atrix) matrices:

Substitution matrices in which scores for each position are derived from *observations* of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity.

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			

BLOSUM 62

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ⟷ *More divergent*

Template Identification (Pairwise comparison)

Because pairwise alignment is slow:

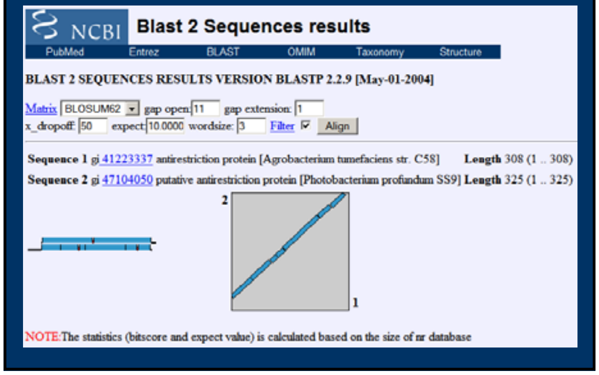
FASTA: Heuristic string alignment program which is based upon finding short exact matches (k-mers) between the target sequence and the database

BLAST: Exploits the speed of exact matching while factoring in the impact of a scoring matrix. Also breaks the query into k-mers, but constructs also k-mers within a certain distance of initial ones.

Conventional wisdom has BLAST as faster than FASTA, but perhaps a little less accurate.

Both are local alignment techniques.

Template Identification (BLAST Example)



Blast 2 Sequences results

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.9 [May-01-2004]

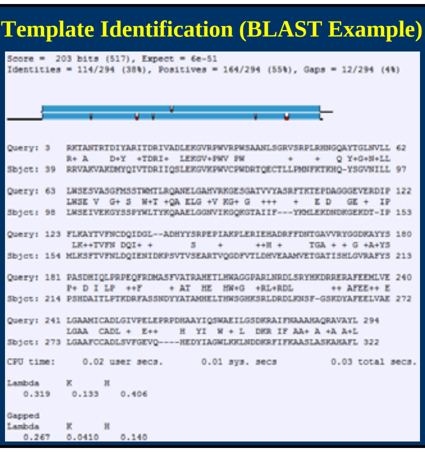
Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.0000 wordsize: 3 Filter ☑ Align

Sequence 1 gi 41223337 antirestriction protein [Agrobacterium tumefaciens str. C58] **Length** 308 (1 .. 308)

Sequence 2 gi 47104050 putative antirestriction protein [Photobacterium profundum SS9] **Length** 325 (1 .. 325)

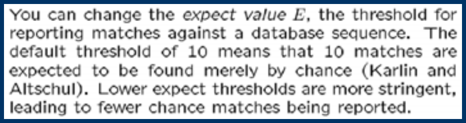
NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

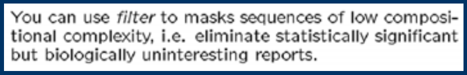
Template Identification (BLAST Example)

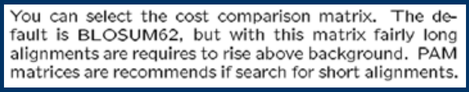


```
Score =  203 bits (517), Expect = 6e-51
Identities = 114/294 (38%), Positives = 164/294 (55%), Gaps = 12/294 (4%)

Query: 3   RKTANTRTDIYARITDRIVADLEKGVRPWVRPWSAANLSGRVSRPLRHNGQAYTGLNVLL 62
            R+ A    D+Y  +TDRI+  LEKGV+PWV PW        +    +   Q Y+G+N+LL
Sbjct: 39  RRVAKVAKDMYQIVTDRIIQSLEKGVKPWVCPWDRTQECTLLPMNFKTKHQ-YSGVNILL 97

Query: 63  LWSESVASGFMSSTWMTLRQANELGAHVRKGESGATVVYASRFTKTEPDAGGGEVERDIP 122
            LWSE V  G+ S  W+T +QA ELG +V KG+ G  +++   +   E D   GE +  IP
Sbjct: 98  LWSEIVEKGYSSPYWLTYKQAAELGGNVIKGQKSTAIIF---YKMLEKDNDKGEKDT-IP 153

Query: 123 FLKAYTVFNCDQIDGL--ADHYYSRPEPIAKPLERIEHADRFFDNTGAVVRYGGDKAYYS 180
             LK++TVFN DQI+ +       S    +      ++H +     TGA + + G +A+YS
Sbjct: 154 MLKSFTVFNLDQIENIDKPSVTVSEARTVQGDFVTLDHVEAAMVETGATISHLGVRAFYS 213

Query: 181 PASDHIQLPRPEQFRDMASFVATRAMETLHWAGGPARLNRDLSRYHKDRRERAFEEMLVE 240
            P+ D I LP  ++F     + AT  HE  HW+G  +RL+RDL       ++ AFEE++ E
Sbjct: 214 PSHDAITLPTKDRFASSNDYYATAMHELTHWSGHKSRLDRDLKNSF-GSKDYAFEELVAE 272

Query: 241 LGAAMICADLGIVPELEPRPDHAAYIQSWAKILGSDKRAIFNAAAHAQRAVAYL 294
            LGAA  CADL +  E++    H  YI  W + L  DKR IF AA+ A +A A+L
Sbjct: 273 LGAAFCCADLSVFGEVQ----HEDYIAGWLKKLNDDKRFIFKAASLASKAHAFL 322

CPU time:     0.02 user secs.       0.01 sys. secs          0.03 total secs.

Lambda     K      H
   0.319    0.133    0.406

Gapped
Lambda     K      H
   0.267    0.0410    0.140
```

Template Identification (BLAST Example)


You can change the *expect value E*, the threshold for reporting matches against a database sequence. The default threshold of 10 means that 10 matches are expected to be found merely by chance (Karlin and Altschul). Lower expect thresholds are more stringent, leading to fewer chance matches being reported.


You can use *filter* to masks sequences of low compositional complexity, i.e. eliminate statistically significant but biologically uninteresting reports.


You can select the cost comparison matrix. The default is BLOSUM62, but with this matrix fairly long alignments are requires to rise above background. PAM matrices are recommends if search for short alignments.

Template Identification (Multiple sequence comparison)

Position Specific Iterative BLAST (PSI-BLAST)

Multiple sequence comparison search, which iteratively expands the set of homologs of the target sequence.

Steps for a given sequence:

1. **Collect initial set of homologs from a sequence database.**
2. **Make weighted multiple alignment of target sequence and homologs.**
3. **Construct position-specific scoring matrix from alignment.**
4. **Use matrix to search database for new homologs.**
5. **Repeat until no new homologs are found.**

Template Identification (Multiple sequence comparison)

PSI-BLAST finds homologs of known structure for approximately twice as many sequences as BLAST.

Variation: Find related sequences in the database for target sequence and create target sequence profile. Find potential templates by comparing the target sequence profile with each of the sequence profiles for known structures.

Variation: Combine multiple alignment information with structural information predicted from the sequence of the target. Especially useful when sequence identity between target and template drops below 25%.

Template Identification (threading)

Pairwise comparison of a target protein sequence and a protein structure

The target sequence is matched against a library of 3D profiles or threaded through a library of 3D folds.

A structure dependent scoring function is used.

Structural constraints on sequence:

- Locally – i.e. secondary structure preferences, Gly/Pro in turns
- Globally – hydrophobic core, residue contacts

Sequence-structure alignment must make sense in 3D:

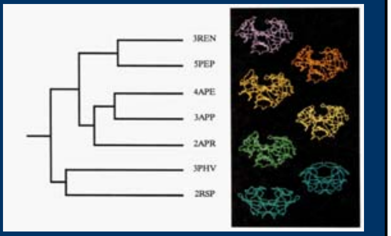
- No gaps in core secondary structures
- No missing strands from sheets

Template Selection

Higher overall sequence similarity between the target and the template sequence usually yields better template.

Other factors:

1. Closest subfamily of proteins, when multiple alignment and phylogenetic tree has been constructed.



Template Selection

2. Template environment considerations, compared to the required environment for the model (e.g. solvent, pH, ligands and quaternay interactions).

3. Quality of the experimental template structure.

Priorities of the criteria should be adjusted depending on the purpose of the comparative model
Example: Protein ligand model => Choice of template with similar ligand probably more important than template resolution.
Example: Analyze geometry of active site => High resolution template most important.

Use of several approximately equidistant templates generally increases model accuracy.

Target-Template alignment

Most accurate alignment possible is imperative.

No current comparative modeling method can recover from a incorrect alignment.

Multiple sequence alignment frequently use program: CLUSTAL

Combinations of previous methods (sequence similarity, profiles, structural information, conservation) can be used.

Model Building

Families of methods:

1. **Modeling by rigid-body assembly**
2. **Modeling by segment matching**
3. **Modeling by satisfaction of spatial restraints**

All models relatively similar when used optimally.

What a method should provide:

1. Permit easy recalculations when changes are made in the alignment.
2. Straightforward to calculate based on several templates
3. Provide tools to incorporate prior knowledge about the target (experimental data or predicted features)

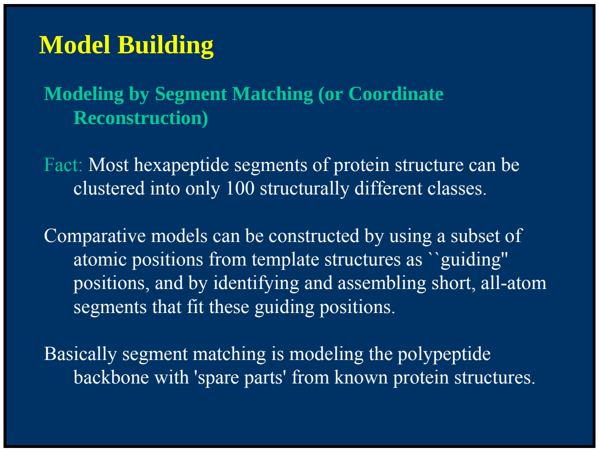
Model Building

Modeling by rigid-body assembly

Assembles a model from a small number of rigid bodies obtained from aligned protein structures. Based on natural dissection of protein folds into conserved core regions, variable loops and side chains.

Steps:

- Template structures selected and superposed
- Frame-work calculation by averaging coordinates of C_α atoms of structurally conserved regions.
- Chain atoms of each core region are generated by comparison of the target with the most similar sequence on the frame-work for this region.
- Loops are generated by scanning whole database of known structures to identify variable regions with similar anchor core regions and compatible sequence.
- Side chains are modeled based on conformational preferences and equivalent template side chain conformations.
- Stereochemistry of the model is improved by restrained energy minimization of MD refinement.



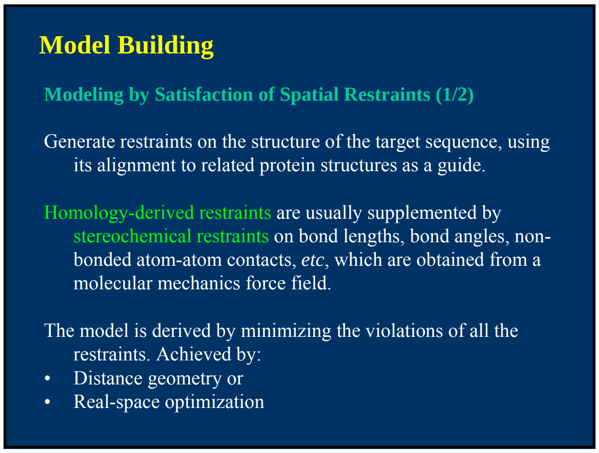
Model Building

Modeling by Segment Matching (or Coordinate Reconstruction)

Fact: Most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes.

Comparative models can be constructed by using a subset of atomic positions from template structures as ``guiding'' positions, and by identifying and assembling short, all-atom segments that fit these guiding positions.

Basically segment matching is modeling the polypeptide backbone with 'spare parts' from known protein structures.



Model Building

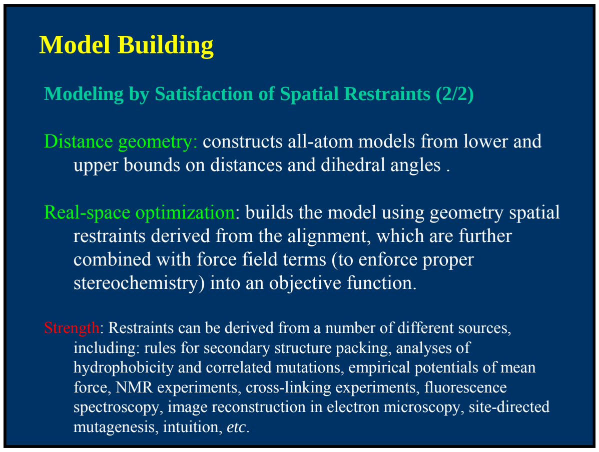
Modeling by Satisfaction of Spatial Restraints (1/2)

Generate restraints on the structure of the target sequence, using its alignment to related protein structures as a guide.

Homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, non-bonded atom-atom contacts, *etc*, which are obtained from a molecular mechanics force field.

The model is derived by minimizing the violations of all the restraints. Achieved by:

- Distance geometry or
- Real-space optimization



Model Building

Modeling by Satisfaction of Spatial Restraints (2/2)

Distance geometry: constructs all-atom models from lower and upper bounds on distances and dihedral angles .

Real-space optimization: builds the model using geometry spatial restraints derived from the alignment, which are further combined with force field terms (to enforce proper stereochemistry) into an objective function.

Strength: Restraints can be derived from a number of different sources, including: rules for secondary structure packing, analyses of hydrophobicity and correlated mutations, empirical potentials of mean force, NMR experiments, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, intuition, *etc*.

Loop Modeling

Loops often play an important role in defining the functional specificity of a given protein framework, forming the active and binding sites.

No general reliable method is available for constructing long loops, especially since there are even 9-residue loops with unrelated conformations in different proteins.

The database search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops (ex. β–hairpins).

However, the database methods are limited by the fact that the number of possible conformations increases exponentially with the length of a loop.

Database search include stem regions of the loop and the search is performed through many known protein structures, not only homologs of the modeled protein.

Errors in comparative modeling

a. Errors in sidechain packing.

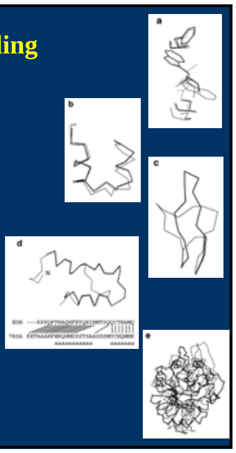
b. Distortions or shifts of a region that is aligned correctly with the template structures.

c. Distortions or shifts of a region that does not have an equivalent segment in any of the template structures.

d. Distortions or shifts of a region that is aligned incorrectly with the template structures.

e. A misfolded structure resulting from using an incorrect template.

Significant methodological improvements are needed to address all of these errors.



Evaluation of models

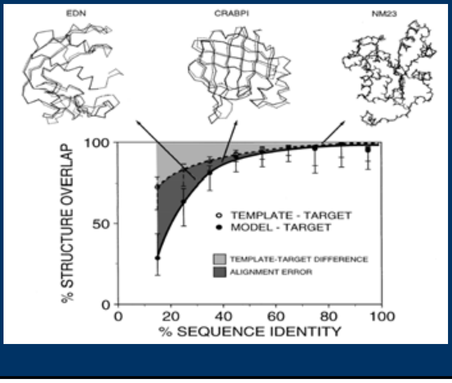
- The 3D models are generally evaluated by relying on geometrical preferences of the amino acid residues or atoms that are derived from known protein structures.

- A basic requirement for a model is to have good stereochemistry. The most useful programs for evaluating stereochemistry are PROCHECK, PROCHECK-NMR, AQUA, SQUID, and WHATCHECK. The features of a model that are checked by these programs include bond lengths, bond angles, peptide bond and sidechain ring planarities, chirality, mainchain and sidechain torsion angles, and clashes between non-bonded pairs of atoms.

- A model also has to have low energy according to a molecular mechanics force field, although low molecular mechanics energy does not ensure that the model is correct.

- Another group of methods for testing 3D models, which implicitly take into account many of the criteria listed above, involve 3D profiles and statistical potentials. Programs implementing this approach include VERIFY3D, PROSA, HARMONY, and ANOLEA.

Evaluation of models

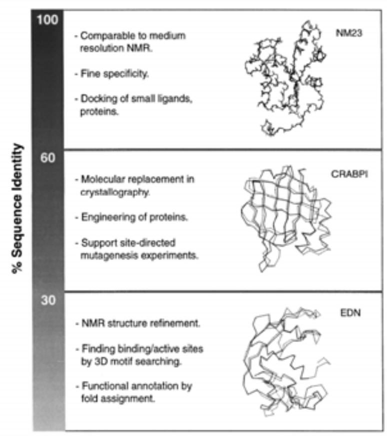


Applications of comparative modeling

Comparative models have been used in:

- studying catalytic mechanisms of enzymes,
- designing and improving ligands,
- docking of macromolecules,
- Predicting interacting protein partners,
- virtual screening and docking of small ligands,
- defining antibody epitopes,
- Molecular replacement in X-ray crystallography,
- designing chimeras, stable and crystallizable variants,
- supporting site-directed mutagenesis,
- refining NMR structures,
- fitting proteins into low-resolution electron density maps,
- finding functional sites by 3D motif searching,
- determining structure from sparse experimental restraints,
- annotating function from structural relationships, and
- finding patches of conserved surface residues.

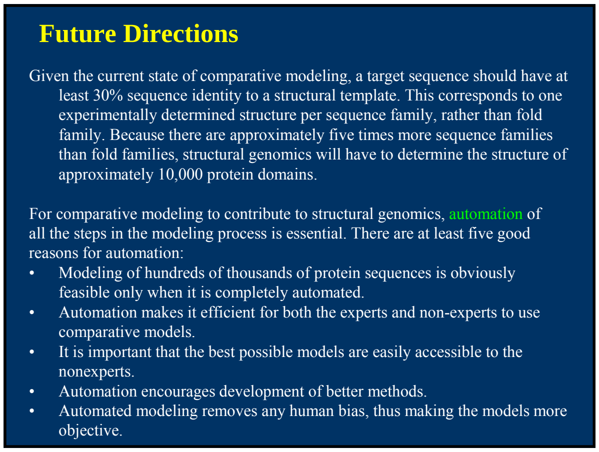
Applications of comparative modeling



- Comparable to medium resolution NMR.
- Fine specificity.
- Docking of small ligands, proteins.

- Molecular replacement in crystallography.
- Engineering of proteins.
- Support site-directed mutagenesis experiments.

- NMR structure refinement.
- Finding binding/active sites by 3D motif searching.
- Functional annotation by fold assignment.

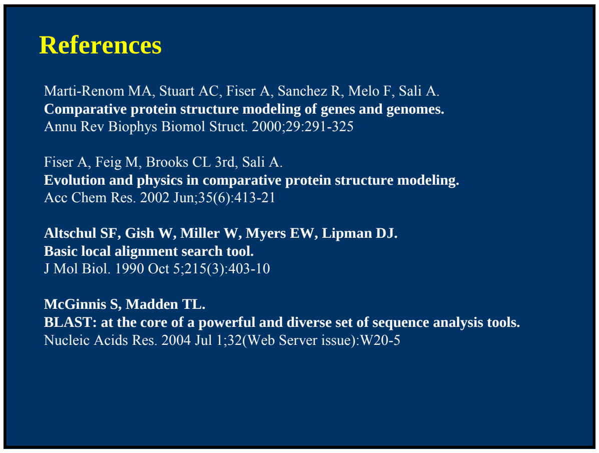


Future Directions

Given the current state of comparative modeling, a target sequence should have at least 30% sequence identity to a structural template. This corresponds to one experimentally determined structure per sequence family, rather than fold family. Because there are approximately five times more sequence families than fold families, structural genomics will have to determine the structure of approximately 10,000 protein domains.

For comparative modeling to contribute to structural genomics, automation of all the steps in the modeling process is essential. There are at least five good reasons for automation:

- Modeling of hundreds of thousands of protein sequences is obviously feasible only when it is completely automated.
- Automation makes it efficient for both the experts and non-experts to use comparative models.
- It is important that the best possible models are easily accessible to the nonexperts.
- Automation encourages development of better methods.
- Automated modeling removes any human bias, thus making the models more objective.



References

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A.
Comparative protein structure modeling of genes and genomes.
Annu Rev Biophys Biomol Struct. 2000;29:291-325

Fiser A, Feig M, Brooks CL 3rd, Sali A.
Evolution and physics in comparative protein structure modeling.
Acc Chem Res. 2002 Jun;35(6):413-21

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
Basic local alignment search tool.
J Mol Biol. 1990 Oct 5;215(3):403-10

McGinnis S, Madden TL.
BLAST: at the core of a powerful and diverse set of sequence analysis tools.
Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W20-5