3D structure prediction

PDB Status (22 March 2017)

Proteins: **127,823**

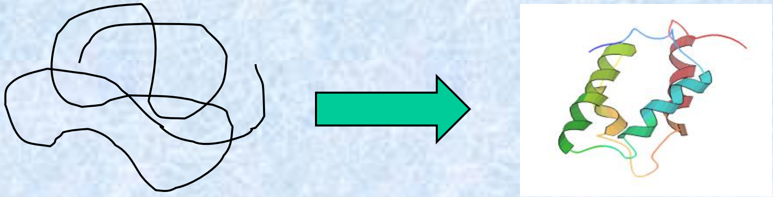
Amino acid sequences are stored in Protein sequence database (SWISS-PROT /EMBL)

Number of available sequences: **~78 million** (increasing exponentially)

Approximate cost
1 structure: $10,000-15,000

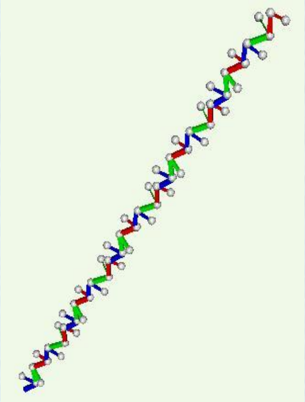
Duration: 2 months – 1 year

Bioinformatics

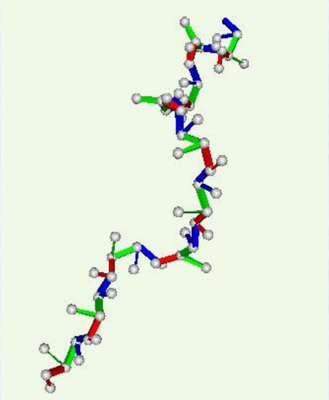


Theme: Deciphering the native conformation of a protein (3D structure) from its amino acid sequence → Protein Folding Problem.

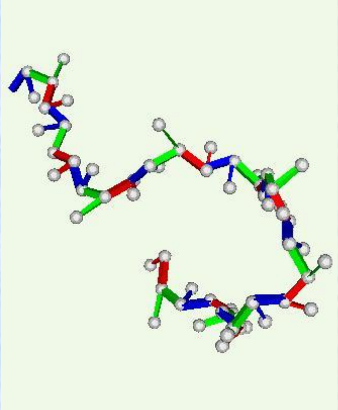
Protein folding problem



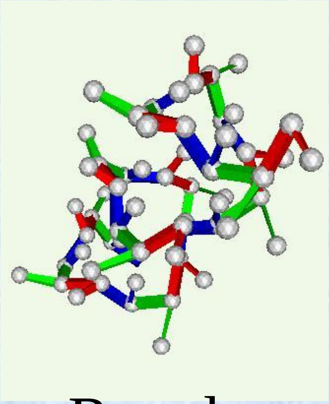
Elongated chain



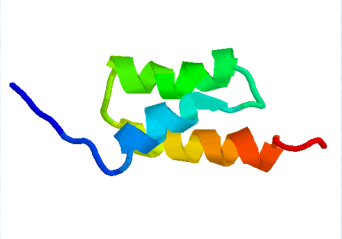
Distortion



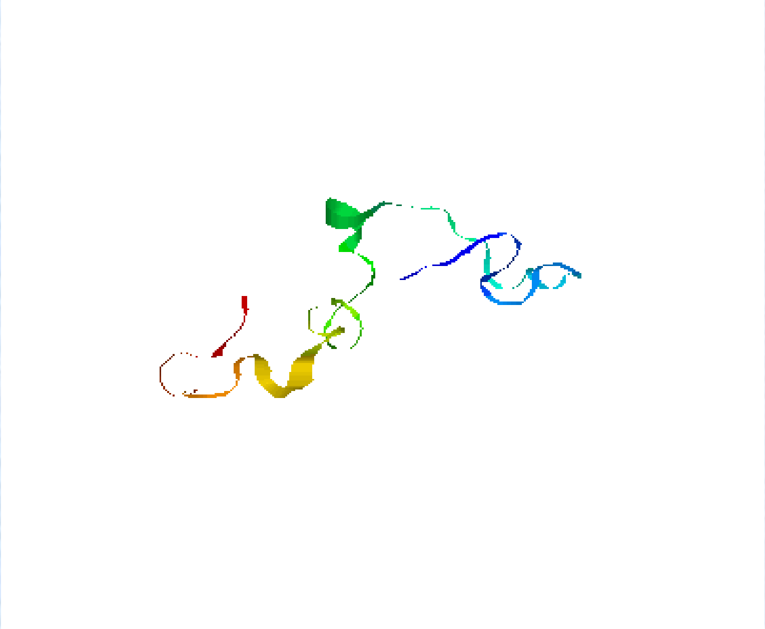
Bends



Regular shape



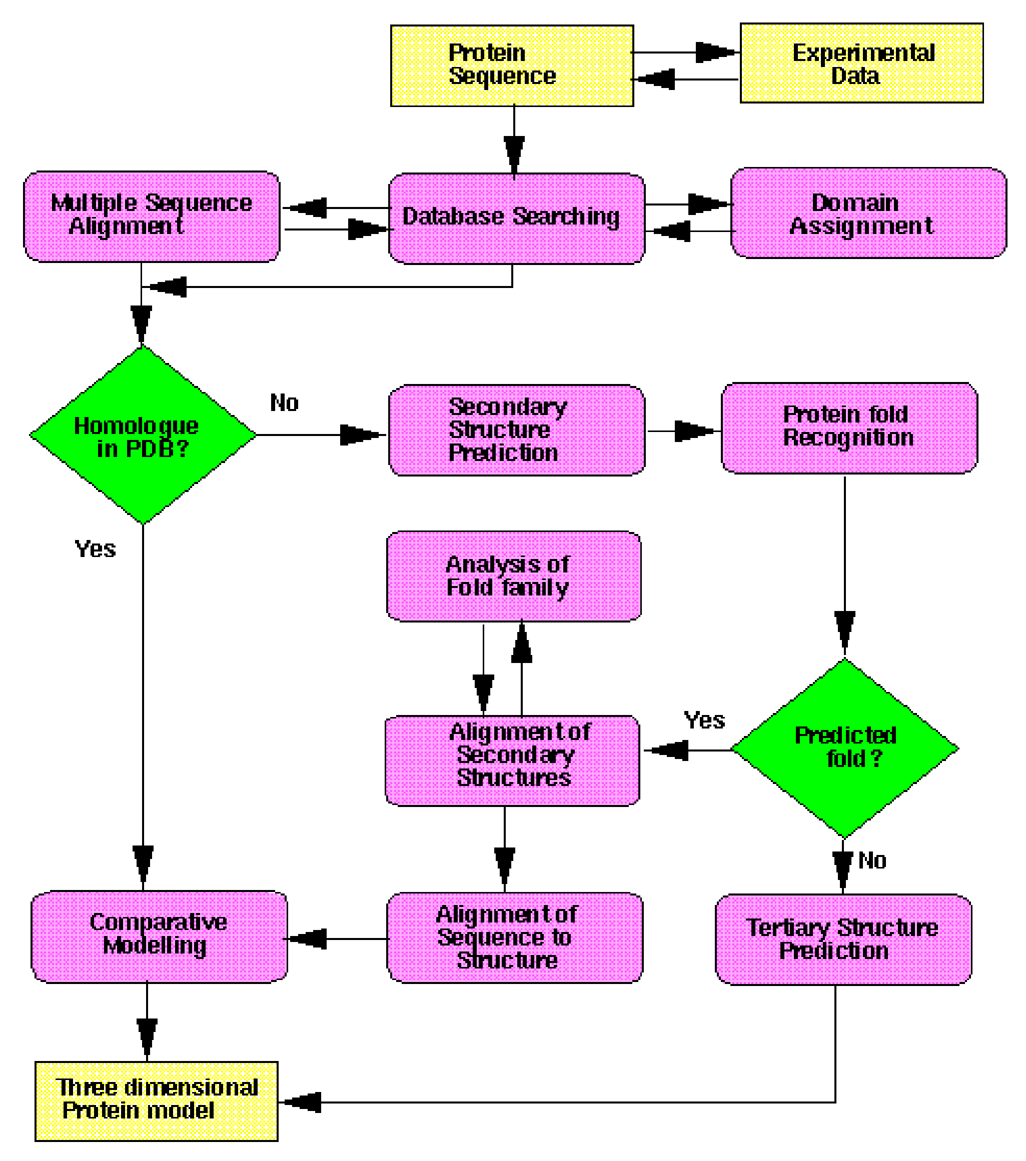
3D structure



3D structure prediction

Critical Assessment of Structure Prediction of Proteins (CASP)

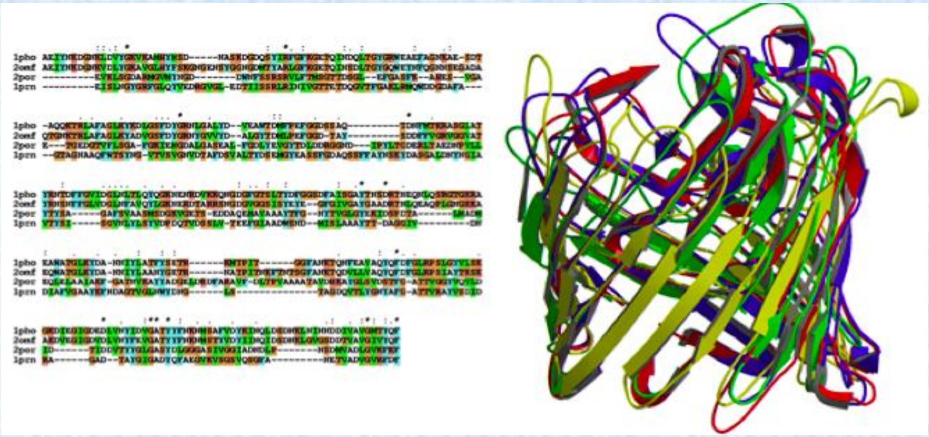
- Homology modeling
- ab initio method (energetic approach)
- Fold recognition



Protein Sequence
Experimental Data
Multiple Sequence Alignment
Database Searching
Domain Assignment
Homologue in PDB?
No
Yes
Secondary Structure Prediction
Protein fold Recognition
Analysis of Fold family
Alignment of Secondary Structures
Yes
Predicted fold ?
No
Comparative Modelling
Alignment of Sequence to Structure
Tertiary Structure Prediction
Three dimensional Protein model

Protein Homology Modelling

Prediction of a 3D structure of a protein from its sequence with an accuracy that is comparable to the best results achieved experimentally.



Allow users to safely use rapidly generated *in silico protein models in all the contexts*

- **Structure-based drug design**
- **Analysis of protein function**
- **Interactions**
- **Antigenic behavior and**
- **Rational design of proteins with increased stability or novel functions.**

Protein modeling is the only way to obtain structural information if experimental techniques fail.

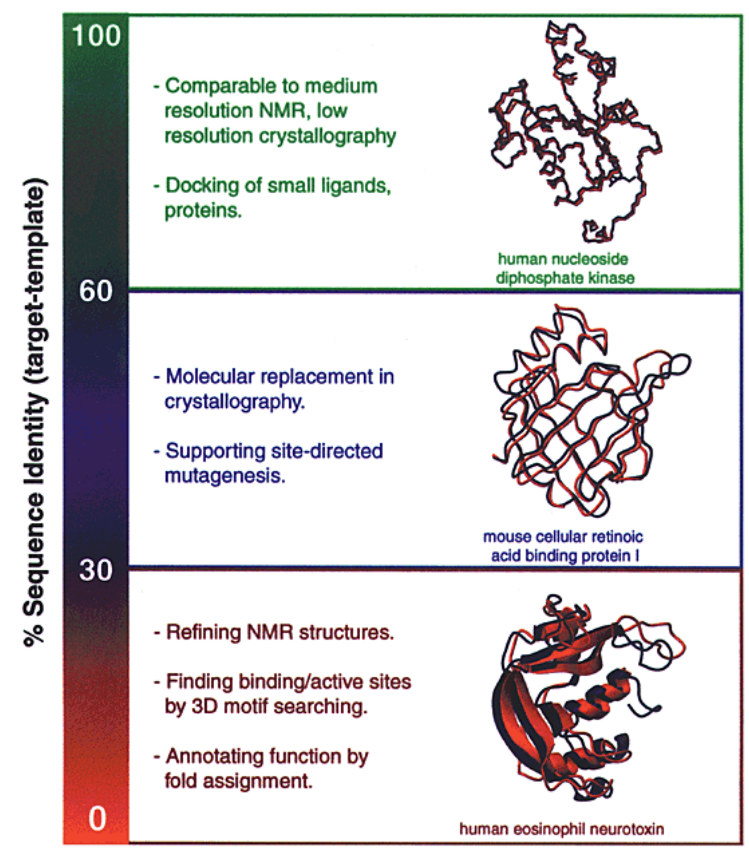
Protein Sequence ➡ Structure

The structure of a protein is uniquely determined by its amino acid sequence (Epstein, Goldberger, and Anfinsen, 1963).

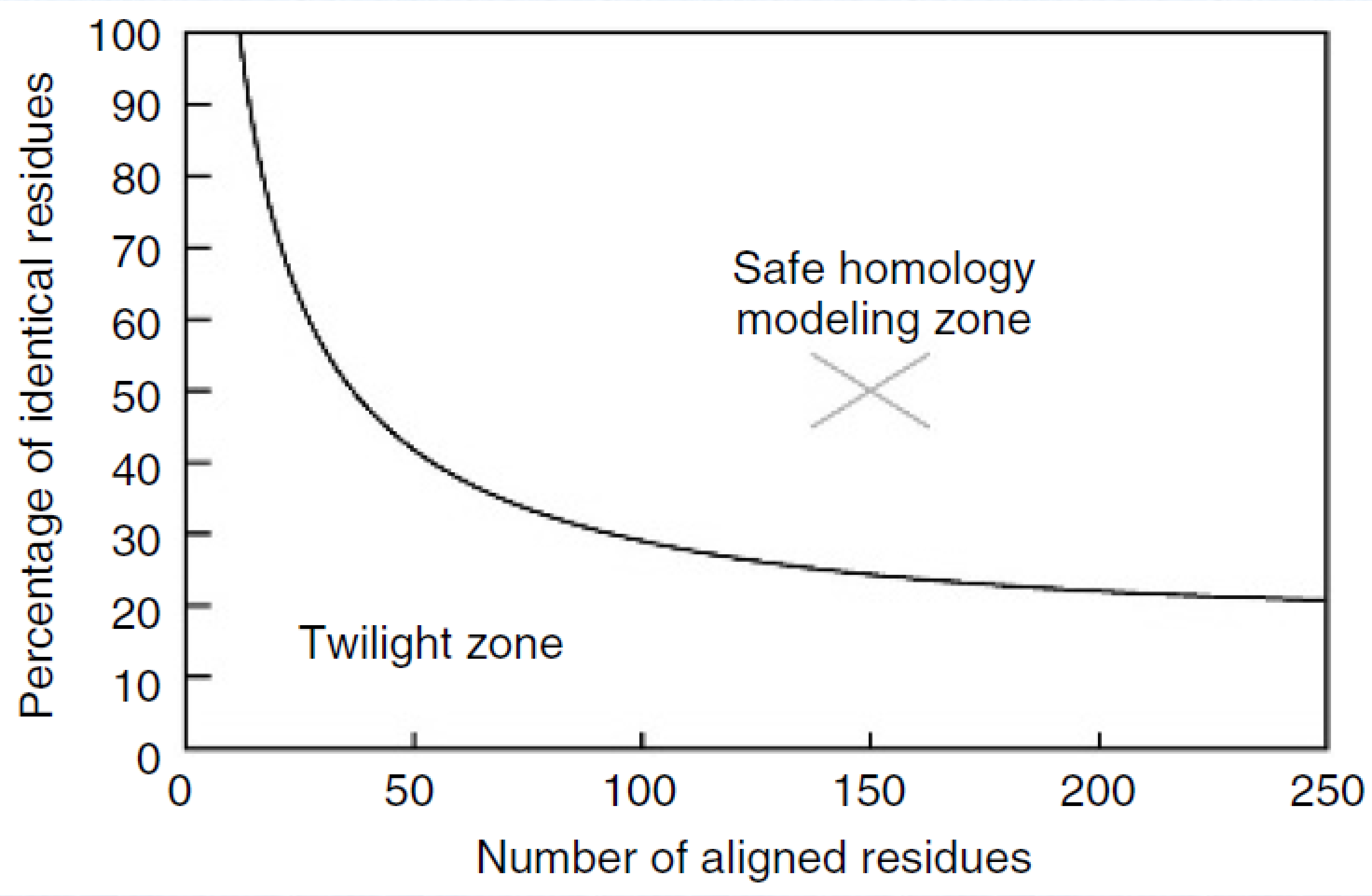
Knowing the sequence should, at least in theory, suffice to obtain the structure.

During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures.

As long as the length of two sequences and the percentage of identical residues fall in the region marked as "safe," the two sequences are practically guaranteed to adopt a similar structure.

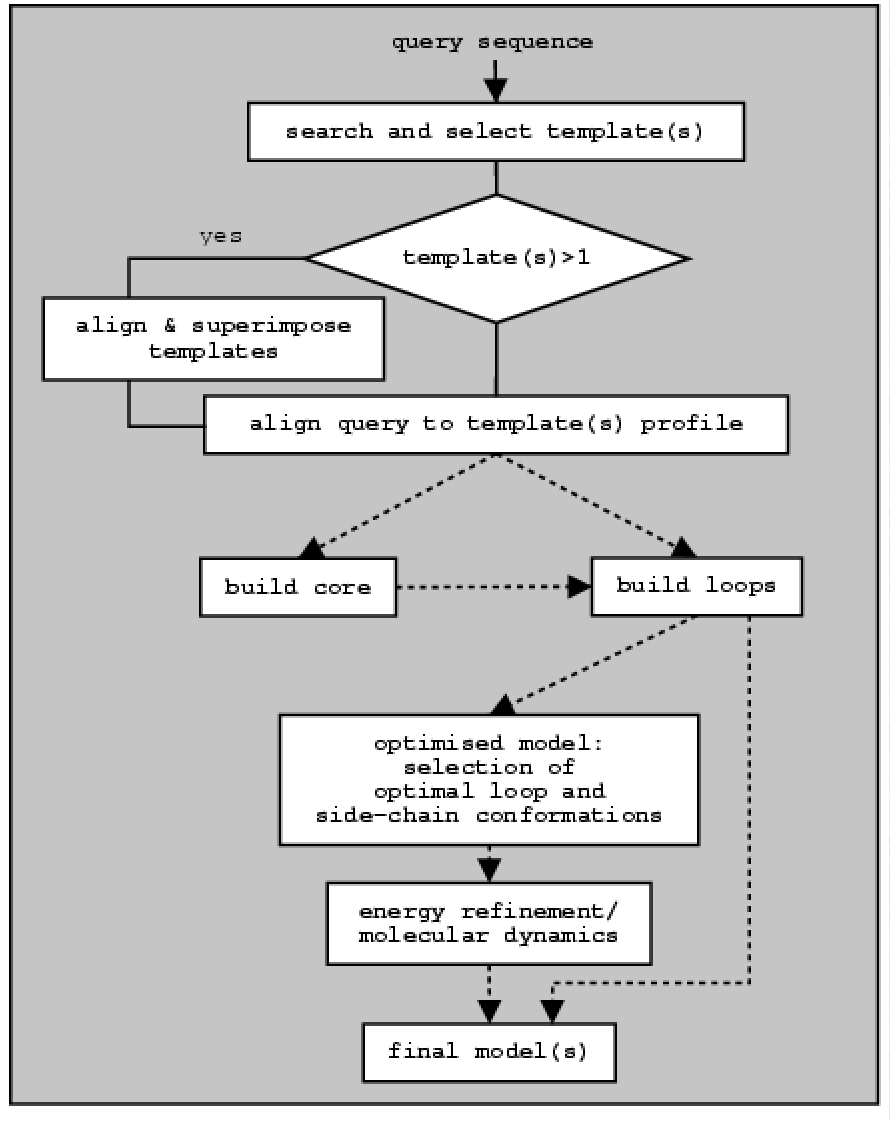

% Sequence Identity (target-template)
100
60
30
0
- Comparable to medium resolution NMR, low resolution crystallography
- Docking of small ligands, proteins.
human nucleoside diphosphate kinase
- Molecular replacement in crystallography.
- Supporting site-directed mutagenesis.
mouse cellular retinoic acid binding protein I
- Refining NMR structures.
- Finding binding/active sites by 3D motif searching.
- Annotating function by fold assignment.
human eosinophil neurotoxin

Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* **22**, M20–M24 (1999)

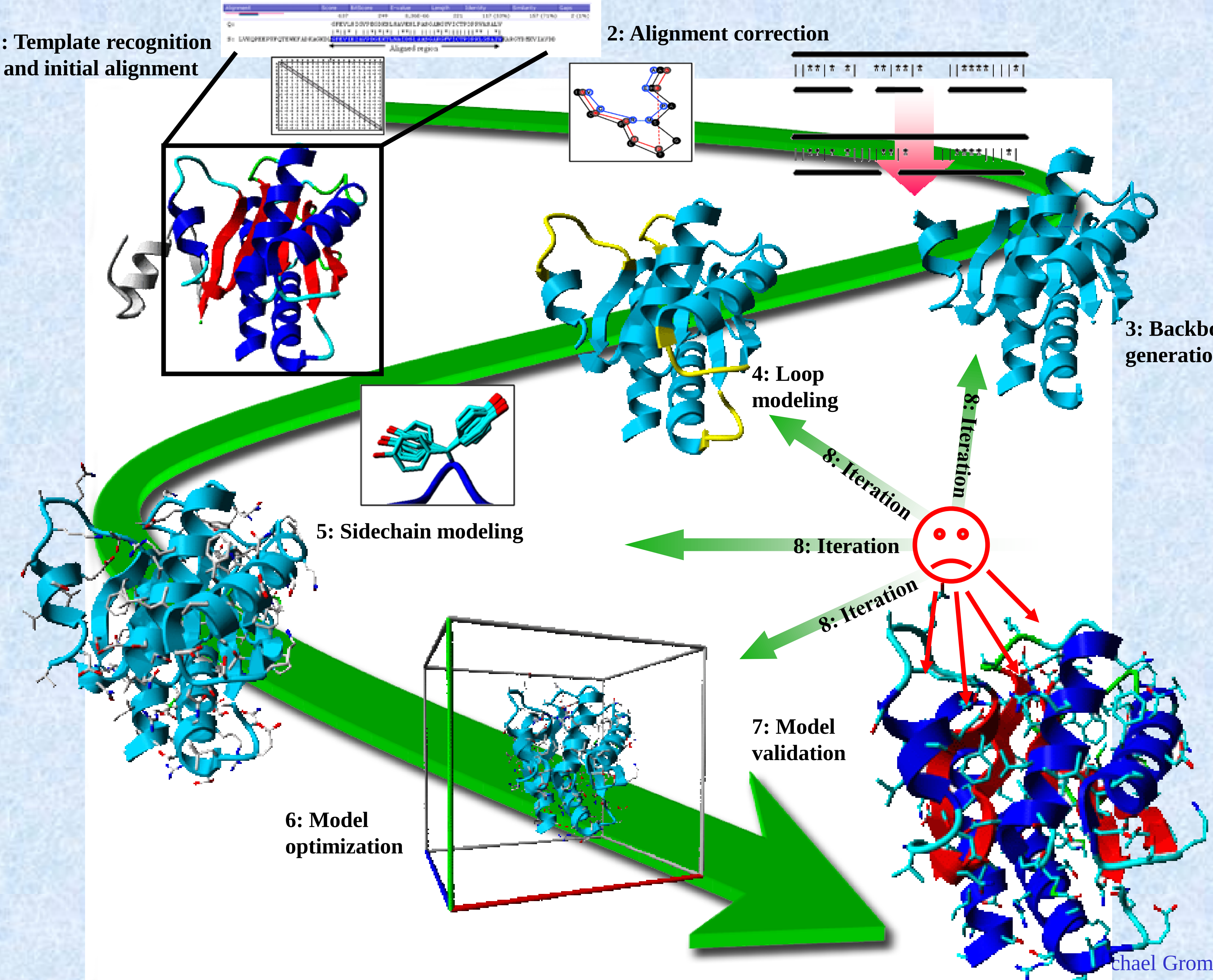


The two zones of sequence alignments. Two sequences are practically guaranteed to fold into the same structure if their length and percentage sequence identity fall into the region marked as "safe." An example of two sequences with 150 amino acids, 50% of which are identical, is shown (gray cross).

Homology Modeling flowchart


query sequence
search and select template(s)
template(s)>1
yes
align & superimpose templates
align query to template(s) profile
build core
build loops
optimised model: selection of optimal loop and side-chain conformations
energy refinement/ molecular dynamics
final model(s)

1. **Template recognition and initial alignment**
2. **Alignment correction**
3. **Backbone generation**
4. **Loop modeling**
5. **Side-chain modeling**
6. **Model optimization**
7. **Model validation**


1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Aligned region

1. Template recognition and initial alignment

- Search the related protein sequences (templates) to the target sequence in any structural database of proteins

- The accuracy of model depends on the selection of proper template

- FASTA and BLAST from EMBL-EBI and NCBI can be used

- Gives a probable set of templates but the final one is not yet decided

- After intial aligments and finding structurally conserved regions among templates, we choose the final template

- Search with sequence
  - Blast
  - Psi-Blast
  - Fold recognition methods

- Use biological information

- Functional annotation in databases

- Active site/motifs

1. Template recognition and initial alignment

- When two or more reference protein structures are available
- Establish structural guidelines for the family of proteins under consideration
- Regions, which are structurally conserved or constant among all the reference proteins
- Target protein is supposed to assume the same conformation in conserved regions

Structurally Conserved Regions

- ❑ SCRs are region in all proteins of a particular family that are nearly identical in structures.
- ❑ Tend to be at inner cores of the proteins
- ❑ Usually contains alpha-helices and beta sheets
- ❑ No SCR can span more than one secondary structure

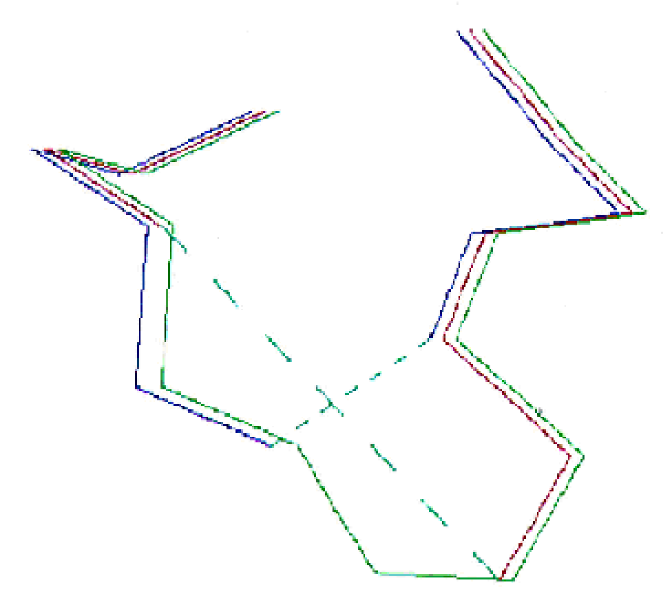
There are generally two main approaches

Constructing c-alpha distance matrix

Aligning vectors of secondary structure units

2. Alignment correction

- It is possible that the alignment has to be corrected.
- A change of Ala to Glu is possible but unlikely to happen in a hydrophobic core, so this Ala and Glu cannot be aligned.
- Using a multiple sequence alignment program (ClustalW) the residues and properties that have to be conserved can be found.
- By looking at the template structure it will become clear which residues are in the core and are less likely to be changed than the residues at the outside.
- Insertions and deletions can be made in widely divergent parts of the molecule and a multiple sequence alignment can be helpful to find these places.
- Gaps have to be shifted around until they are as small as possible.



Template structure (green) with the best aligned target (red) with a large gap, and the target after shifting several residues (blue). The gap is much smaller now.

How to find an appropriate template Structure for homology modelling…

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

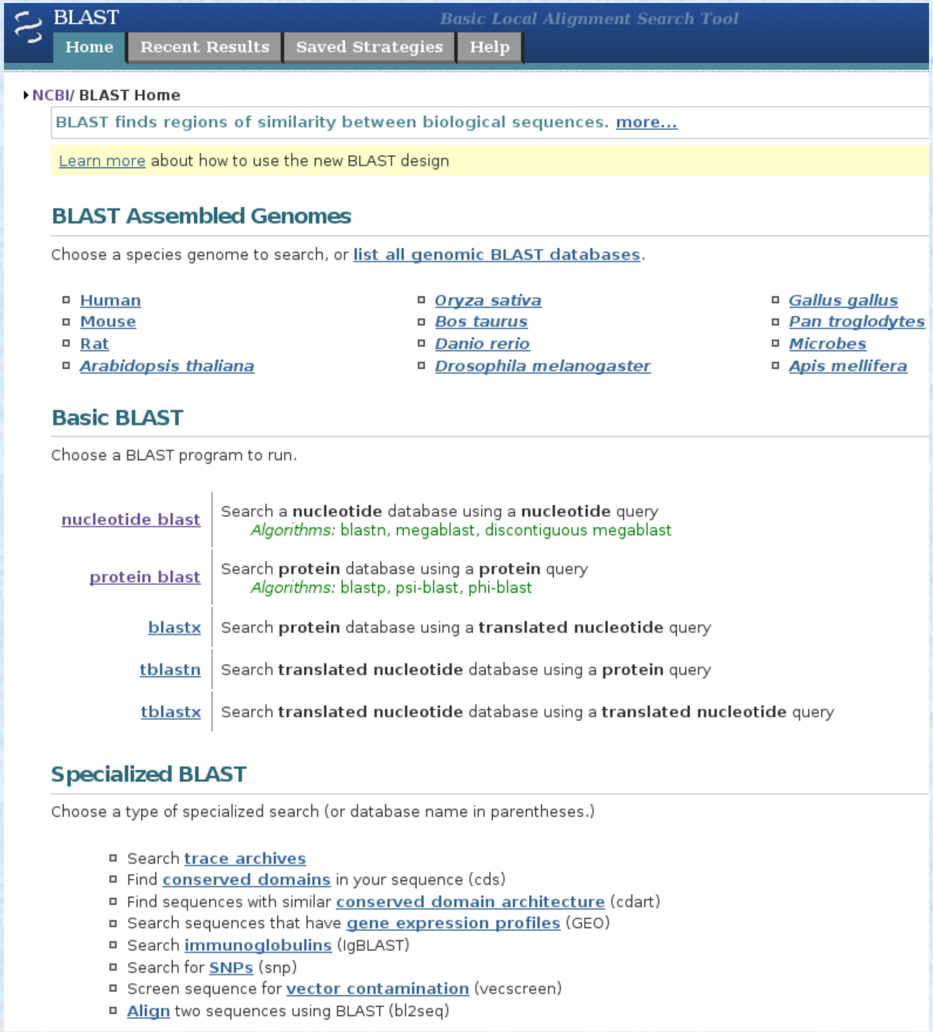
- Used to search protein databases:

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

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

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

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



NCBI/ BLAST Home

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

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

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

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

Basic BLAST

Choose a BLAST program to run.

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

Specialized BLAST

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

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

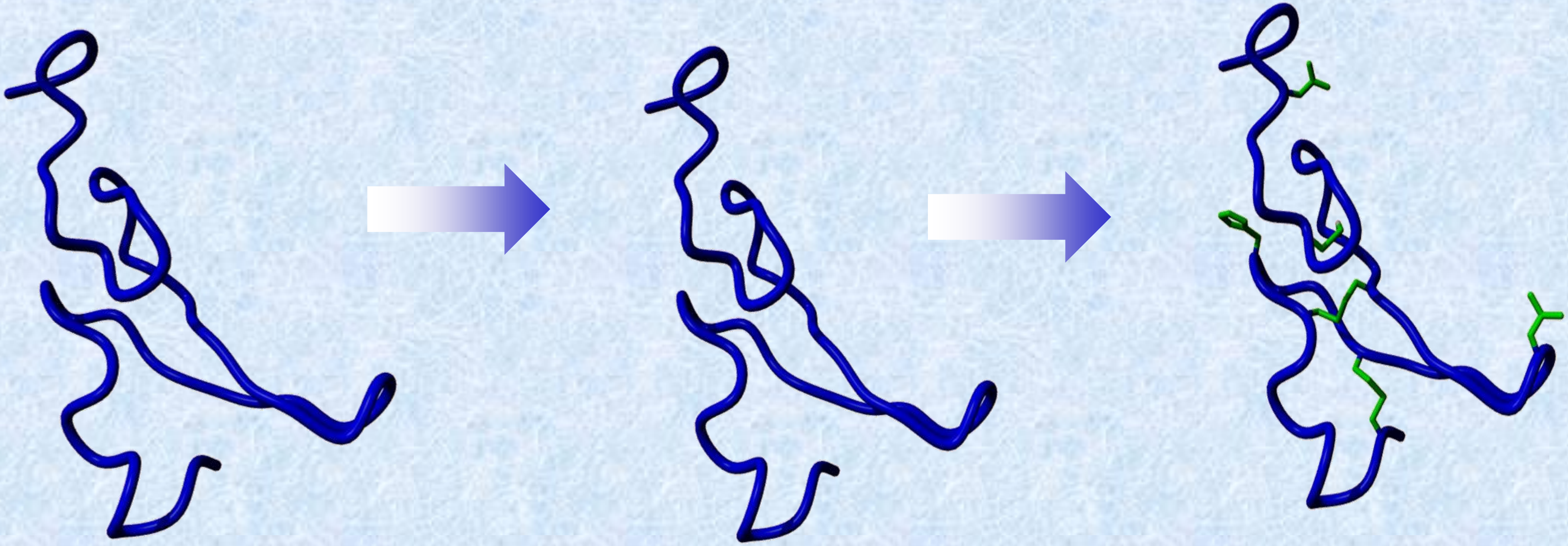
3. Backbone generation

When the alignment is correct, the backbone of the target can be created.

The coordinates of the template-backbone are copied to the target.

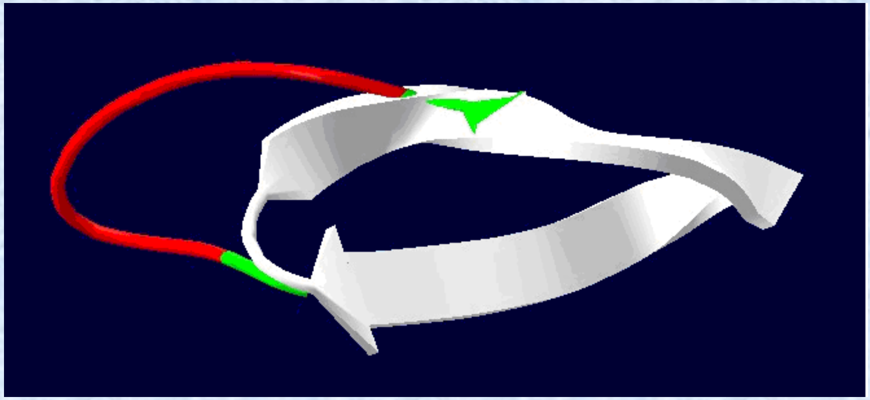
When the residues are identical, the side-chain coordinates are also copied.

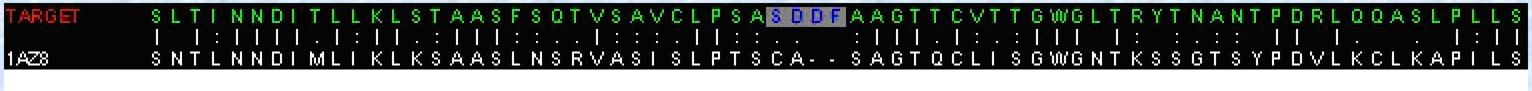
Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.



4. Loop modeling

- ❑ Check the loops on the basis of steric overlaps
- ❑ A specified degree of overlap can be tolerated
- ❑ Check the atoms within the loop agains each other
- ❑ Then check loop atoms against rest of the protein’s atoms





```
TARGET   S L T I N N D I T L L K L S T A A S F S Q T V S A V C L P S A S D D F A A G T T C V T T G W G L T R Y T N A N T P D R L Q Q A S L P L L S
         |   | : | | | | . | : | | . : | | | : : . . | : : : . | | : : . .     : | | | . | : . : | | |   | :   : . : :   | |   | .     .   | : | |
1AZ8     S N T L N N D I M L I K L K S A A S L N S R V A S I S L P T S C A - - S A G T Q C L I S G W G N T K S S G T S Y P D V L K C L K A P I L S
```

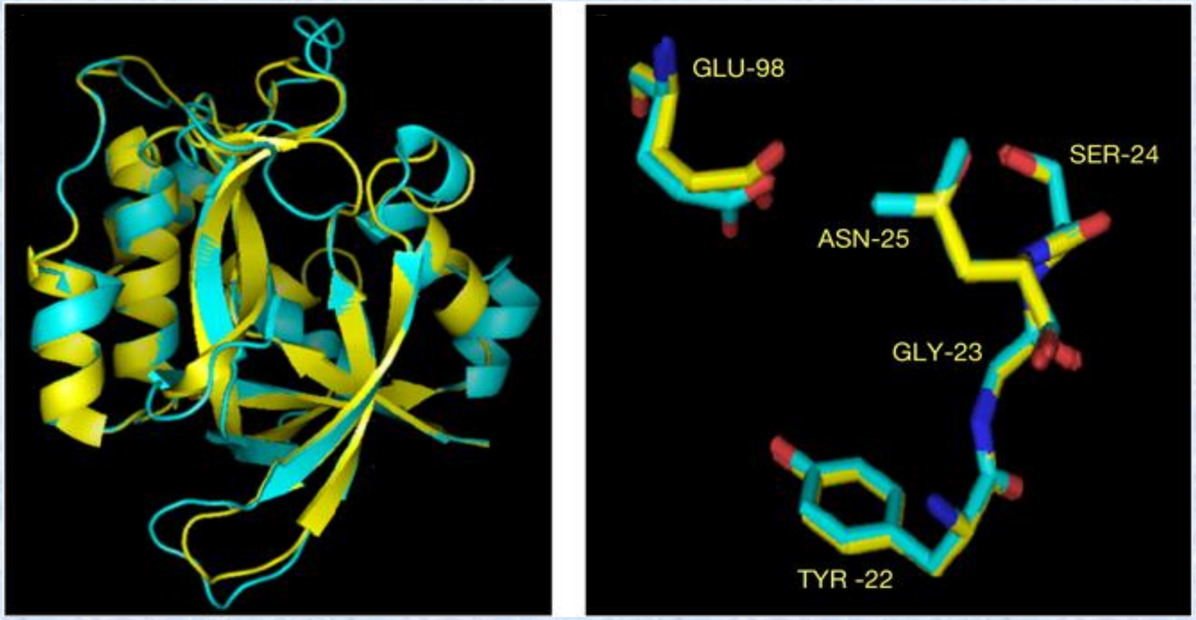
5. Side-chain modeling

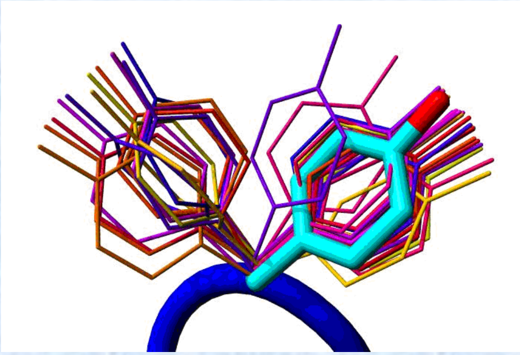
- ❑ With bond lengths, bond angles and two rotable backbone bonds per residue φ and ψ, its very difficult to find the best conformation of a side chain

- ❑ In addition, side chains of many residues have one or more degree of freedom.

- ❑ Side chain conformational search in loop regions is must

- ❑ Side chain residues replaced during coordinate transformations should also be checked





Rotamer

- Statstical studies show side chain adopt only a small number of many possible conformations
- The correct rotamer of a particular residue is mainly determined by local environment
- Side chain generally adopt conformations where they are closely packed

Observations:

- In homologous proteins, corresponding residues virtually retain the same rotameric state
- Within a range of χ values, 80% of the identical residues and 75% of the mutated residues have the same conformations
- Certain rotamers are almost always associated with certain secondary structure

6. Model optimization

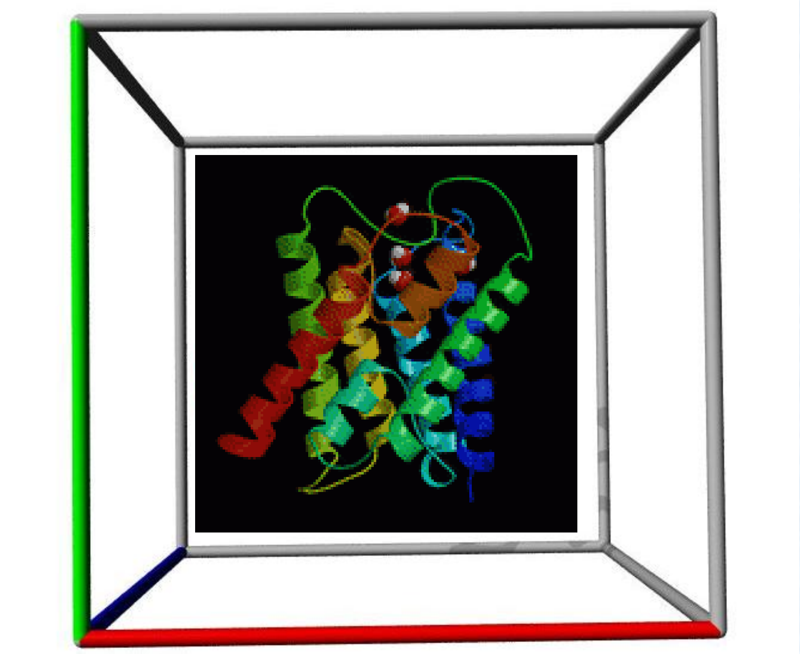
Many **structural artifacts** can be introduced while the model protein is being built

- ❑ Substitution of large side chains for small ones
- ❑ Strained peptide bonds between segments taken from difference reference proteins
- ❑ Non optimum conformation of loops

Molecular dynamics simulation

Remove big errors

Structure moves to lowest energy conformation



❑ Energy Minimisation is used to produce a chemically and conformationally reasonable model protein structure

❑ Two mainly used optimisation algorithms are

- ➢ Steepest Descent
- ➢ Conjugate Gradients

❑ Molecular Dynamics is used to explore the conformational space a molecule could visit

7. Model validation

- **Every homology model contains errors.Two main reasons**
  - **% sequence identity between reference and model**
  - **The number of errors in templates**
- **Hence it is essential to check the correctness of overall fold/ structure, errors of localized regions and stereochemical parameters: bond lengths, angles, geometries**

  - WHAT IF http://www.cmbi.kun.nl/gv/servers/WIWWWI/
  - SOV http://predictioncenter.llnl.gov/local/sov/sov.html
  - PROVE http://www.ucmb.ulb.ac.be/UCMB/PROVE/
  - ANOLEA http://www.fundp.ac.be/pub/ANOLEA.html
  - ERRAT http://www.doe-mbi.ucla.edu/Services/ERRATv2/
  - VERIFY3D http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D/
  - BIOTECH http://biotech.embl-ebi.ac.uk:8400/
  - ProsaII http://www.came.sbg.ac.at
  - WHATCHECK http://www.sander.embl-heidelberg.de/whatcheck/

Ramachandran Plot

- The results of the Ramachandran plot will be very similar to that of the template.

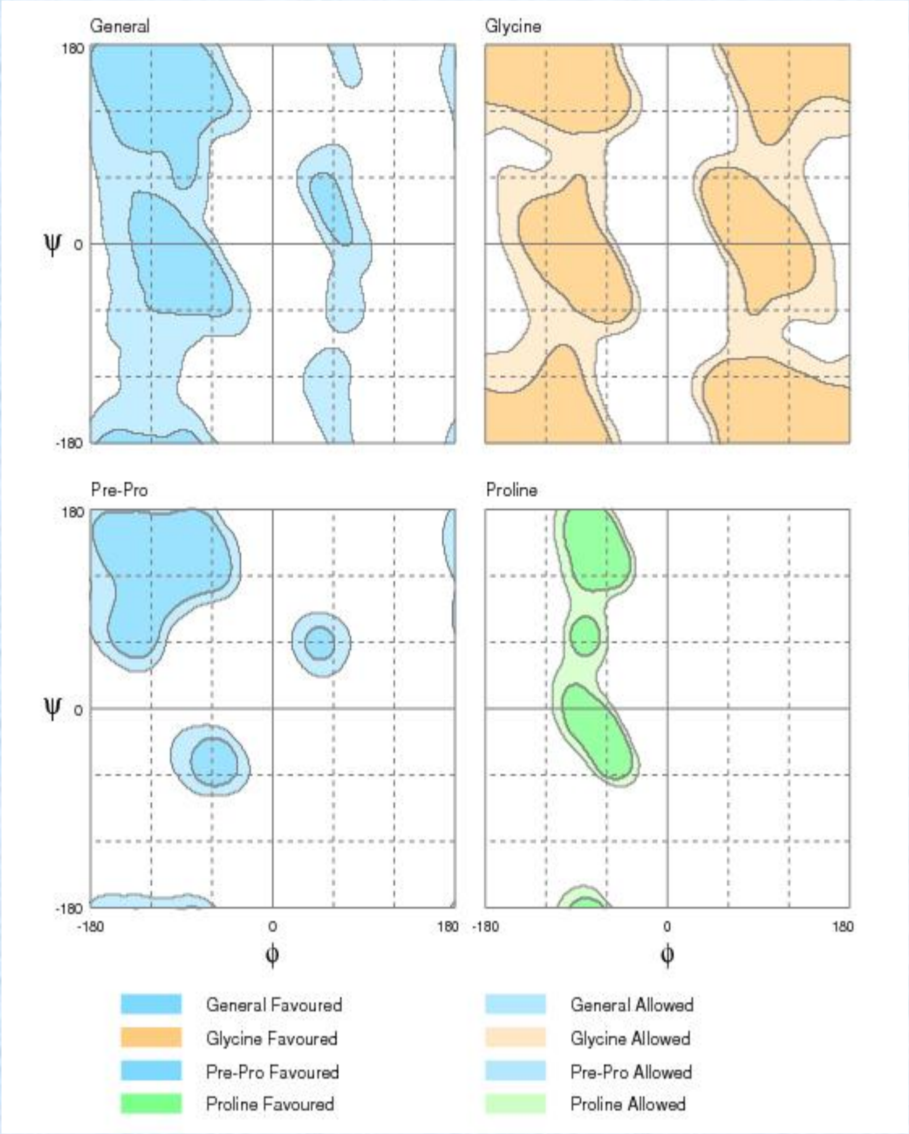
- A **Good template** is therefore key!

- Most residues are mainly found on the **left-hand side of the plot.**

- **Glycine** is found more randomly within plot (orange), due to its small sidechain (H) preventing clashes with its backbone.

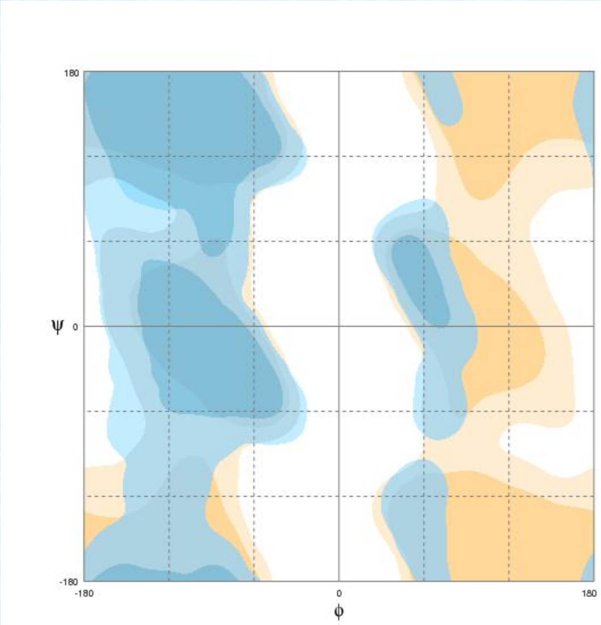
- **Proline** can only adopt a **Phi angle of ~-60°** (green) due to its sidechain.

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



More Advanced Options

- **Procheck, PROVE, WhatIf:**
  Stereochemical checks on bond lengths, angles and atomic contacts.
- **Ramachandran Plot** is a major component of the evaluation.
- Ensures that the backbone conformation of the model is normal.
- Modeller is good on the whole, but sometimes struggles with residues found in loops.

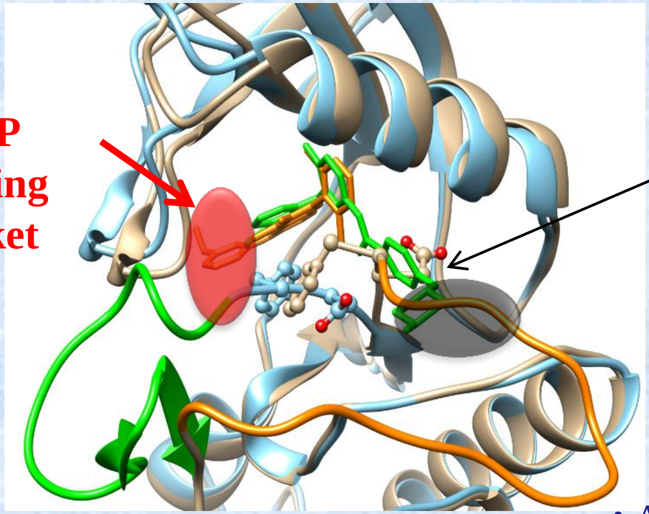


```
 +----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
 |                                                                            |
 | mgirk .pdb   2.5                                              104 residues |
 |                                                      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

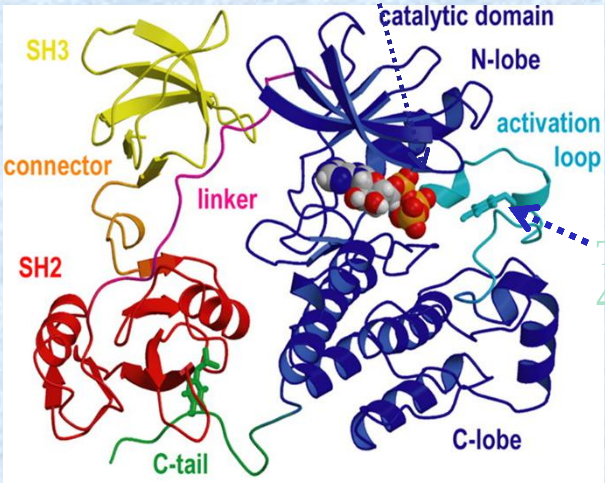
Case Study

```
UniProtKB - P07947 (YES_HUMAN)
>sp|P07947|YES_HUMAN Tyrosine-protein kinase Yes OS=Homo sapiens GN=YES1 PE=1 SV=3
MGCIKSKENKSPAIKYRPENTPEPVSTSVSHYGAEPTTVSPCPSSSAKGTAVNFSSLSMT
PFGGSSGVTPFGGASSSFSVVPSSYPAGLTGGVTIFVALYDYEARTTEDLSFKKGERFQI
INNTEGDWWEARSIATGKNGYIPSNYVAPADSIQAEEWYFGKMGRKDAERLLLNPGNQRG
IFLVRESETTKGAYSLSIRDWDEIRGDNVKHYKIRKLDNGGYYITTRAQFDTLQKLVKHY
TEHADGLCHKLTTVCPTVKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTT
KVAIKTLKPGTMMPEAFLQEAQIMKKLRHDKLVPLYAVVSEEPIYIVTEFMSKGSLLDFL
KEGDGKYLKLPQLVDMAAQIADGMAYIERMNYIHRDLRAANILVGENLVCKIADFGLARL
IEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILQTELVTKGRVPYPGMVNRE
VLEQVERGYRMPCPQGCPESLHELMNLCWKKDPDERPTFEYIQSFLEDYFTATEPQYQPG
ENL
```

ATP binding pocket



DFG Motif (Ball and stick)

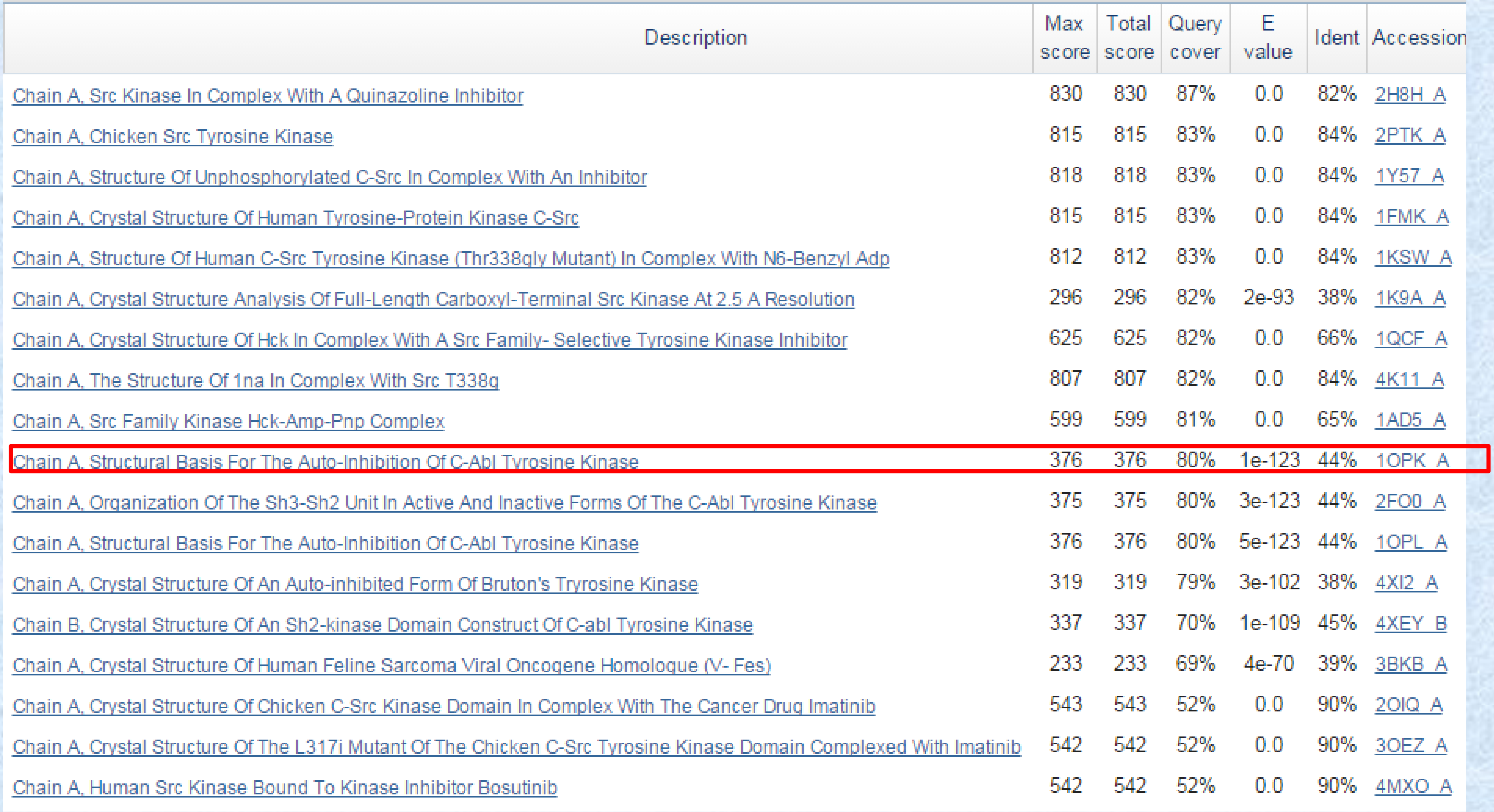


Tyr 416

ABL1 kinase in complex with Imatinib
Type 2 inhibitor
DFG-OUT
Template PDB id: 1IEP

- ABL1 kinase in complex with PD166326
- Type 1 inhibitors
- ATP Competitive
- **DFG-IN**
- Template PDB id: 1OPK

Searching templates for DFG-IN and DFG-OUT conformations

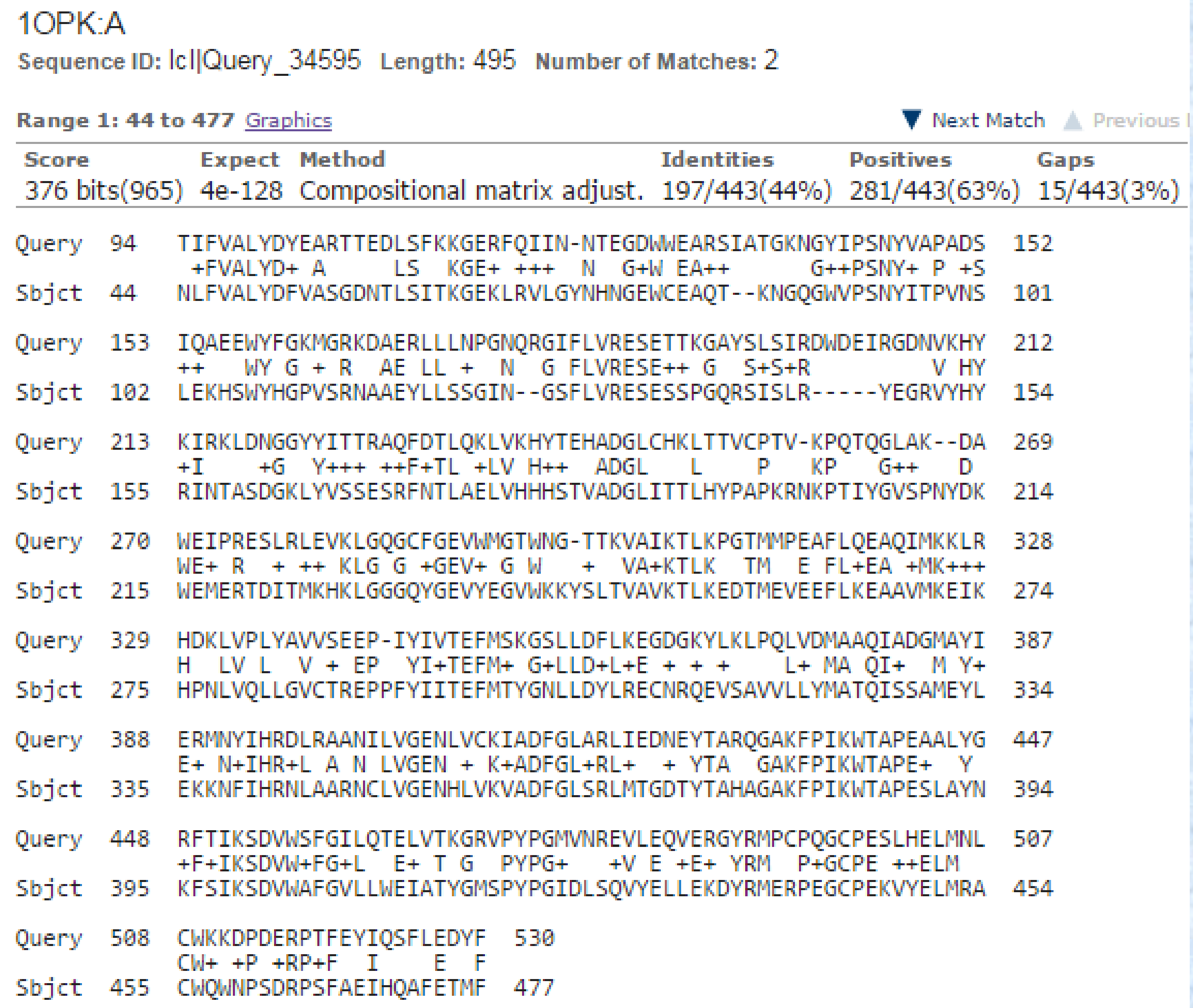


Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Src Kinase In Complex With A Quinazoline Inhibitor	830	830	87%	0.0	82%	2H8H_A
Chain A, Chicken Src Tyrosine Kinase	815	815	83%	0.0	84%	2PTK_A
Chain A, Structure Of Unphosphorylated C-Src In Complex With An Inhibitor	818	818	83%	0.0	84%	1Y57_A
Chain A, Crystal Structure Of Human Tyrosine-Protein Kinase C-Src	815	815	83%	0.0	84%	1FMK_A
Chain A, Structure Of Human C-Src Tyrosine Kinase (Thr338gly Mutant) In Complex With N6-Benzyl Adp	812	812	83%	0.0	84%	1KSW_A
Chain A, Crystal Structure Analysis Of Full-Length Carboxyl-Terminal Src Kinase At 2.5 A Resolution	296	296	82%	2e-93	38%	1K9A_A
Chain A, Crystal Structure Of Hck In Complex With A Src Family- Selective Tyrosine Kinase Inhibitor	625	625	82%	0.0	66%	1QCF_A
Chain A, The Structure Of 1na In Complex With Src T338g	807	807	82%	0.0	84%	4K11_A
Chain A, Src Family Kinase Hck-Amp-Pnp Complex	599	599	81%	0.0	65%	1AD5_A
Chain A, Structural Basis For The Auto-Inhibition Of C-Abl Tyrosine Kinase	376	376	80%	1e-123	44%	1OPK_A
Chain A, Organization Of The Sh3-Sh2 Unit In Active And Inactive Forms Of The C-Abl Tyrosine Kinase	375	375	80%	3e-123	44%	2FO0_A
Chain A, Structural Basis For The Auto-Inhibition Of C-Abl Tyrosine Kinase	376	376	80%	5e-123	44%	1OPL_A
Chain A, Crystal Structure Of An Auto-inhibited Form Of Bruton's Tryrosine Kinase	319	319	79%	3e-102	38%	4XI2_A
Chain B, Crystal Structure Of An Sh2-kinase Domain Construct Of C-abl Tyrosine Kinase	337	337	70%	1e-109	45%	4XEY_B
Chain A, Crystal Structure Of Human Feline Sarcoma Viral Oncogene Homologue (V- Fes)	233	233	69%	4e-70	39%	3BKB_A
Chain A, Crystal Structure Of Chicken C-Src Kinase Domain In Complex With The Cancer Drug Imatinib	543	543	52%	0.0	90%	2OIQ_A
Chain A, Crystal Structure Of The L317i Mutant Of The Chicken C-Src Tyrosine Kinase Domain Complexed With Imatinib	542	542	52%	0.0	90%	3OEZ_A
Chain A, Human Src Kinase Bound To Kinase Inhibitor Bosutinib	542	542	52%	0.0	90%	4MXO_A

- **1OPK is selected as a template to model cYES kinase structure with DFG-IN conformation. It is due to 80% query coverage and well defined active site for binding of type 1 kinase inhibitors.**

- **Moreover, crystal structure of corresponding DFG-OUT conformation (1IEP) is also available and it is used as a template to model cYES kinase structure with DFG-OUT conformation.**

Target – Template Sequence Alignment (BLASTp)



1OPK:A

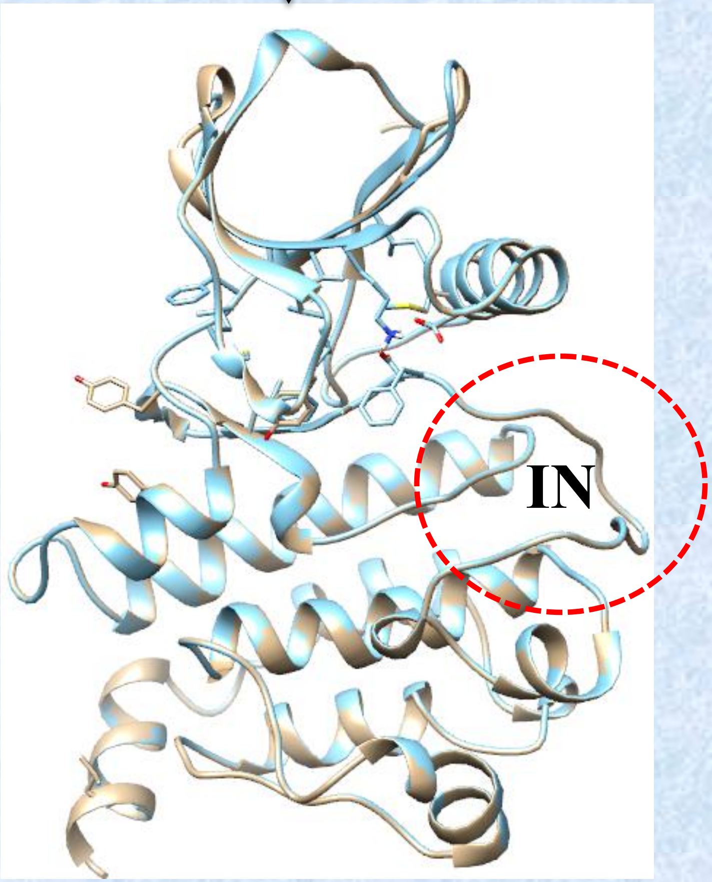
Sequence ID: lcl|Query_34595 Length: 495 Number of Matches: 2

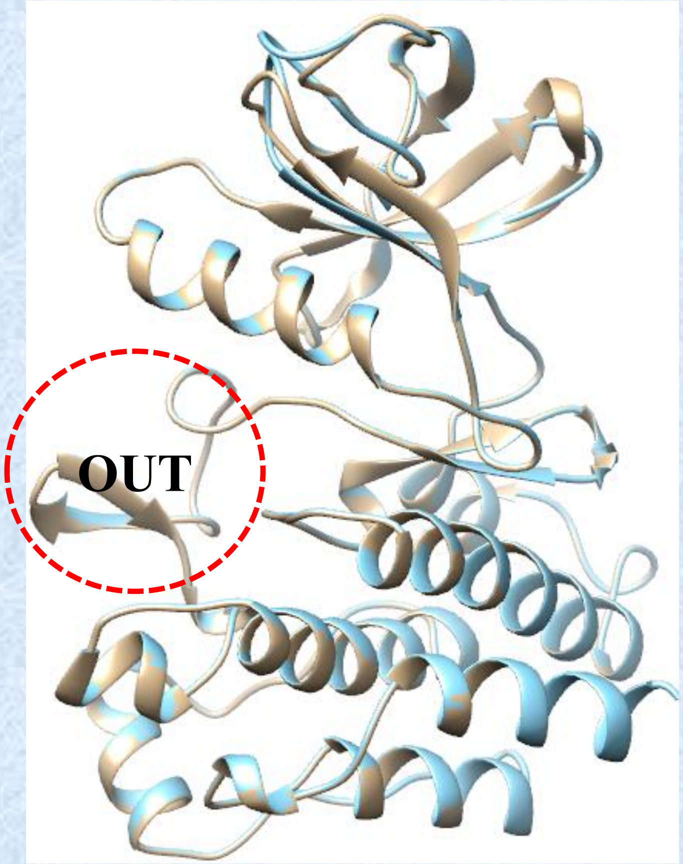
Range 1: 44 to 477 Graphics

Score	Expect	Method	Identities	Positives	Gaps
376 bits(965)	4e-128	Compositional matrix adjust.	197/443(44%)	281/443(63%)	15/443(3%)

```
Query  94    TIFVALYDYEARTTEDLSFKKGERFQIIN-NTEGDWWEARSIATGKNGYIPSNYVAPADS  152
              +FVALYD+ A     LS  KGE+ +++  N  G+W EA++      G++PSNY+ P +S
Sbjct  44    NLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQT--KNGQGWVPSNYITPVNS  101

Query  153   IQAEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETTKGAYSLSIRDWDEIRGDNVKHY  212
             ++   WY G + R  AE LL +  N  G FLVRESE++ G  S+S+R         V HY
Sbjct  102   LEKHSWYHGPVSRNAAEYLLSSGIN--GSFLVRESESSPGQRSISLR-----YEGRVYHY  154

Query  213   KIRKLDNGGYYITTRAQFDTLQKLVKHYTEHADGLCHKLTTVCPTV-KPQTQGLAK--DA  269
             +I    +G  Y+++ ++F+TL +LV H++  ADGL   L    P   KP   G++   D
Sbjct  155   RINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTIYGVSPNYDK  214

Query  270   WEIPRESLRLEVKLGQGCFGEVWMGTWNG-TTKVAIKTLKPGTMMPEAFLQEAQIMKKLR  328
             WE+ R  + ++ KLG G +GEV+ G W   +  VA+KTLK  TM  E FL+EA +MK+++
Sbjct  215   WEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIK  274

Query  329   HDKLVPLYAVVSEEP-IYIVTEFMSKGSLLDFLKEGDGKYLKLPQLVDMAAQIADGMAYI  387
             H  LV L  V + EP  YI+TEFM+ G+LLD+L+E + + +    L+ MA QI+  M Y+
Sbjct  275   HPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVSAVVLLYMATQISSAMEYL  334

Query  388   ERMNYIHRDLRAANILVGENLVCKIADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYG  447
             E+ N+IHR+L A N LVGEN + K+ADFGL+RL+  + YTA  GAKFPIKWTAPE+  Y
Sbjct  335   EKKNFIHRNLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYN  394

Query  448   RFTIKSDVWSFGILQTELVTKGRVPYPGMVNREVLEQVERGYRMPCPQGCPESLHELMNL  507
             +F+IKSDVW+FG+L  E+ T G  PYPG+   +V E +E+ YRM  P+GCPE ++ELM
Sbjct  395   KFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRA  454

Query  508   CWKKDPDERPTFEYIQSFLEDYF  530
             CW+ +P +RP+F  I    E  F
Sbjct  455   CWQWNPSDRPSFAEIHQAFETMF  477
```

Kinases	RMSD	
	Back Bone	All
1OPK vs cYes (DFG-IN)	0.18	0.47
1IEP vs cYes (DFG-OUT)	0.39	0.67

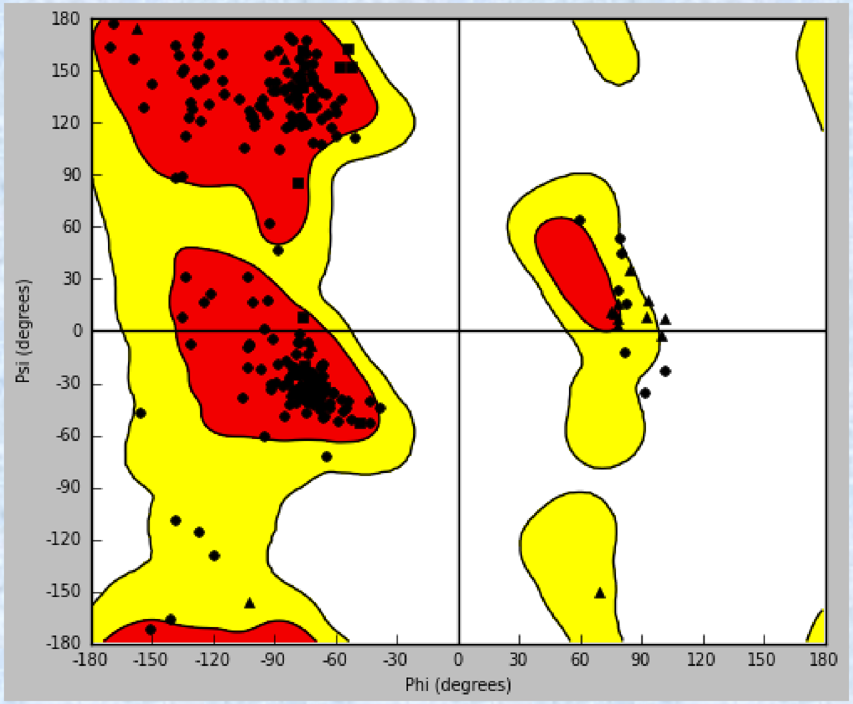


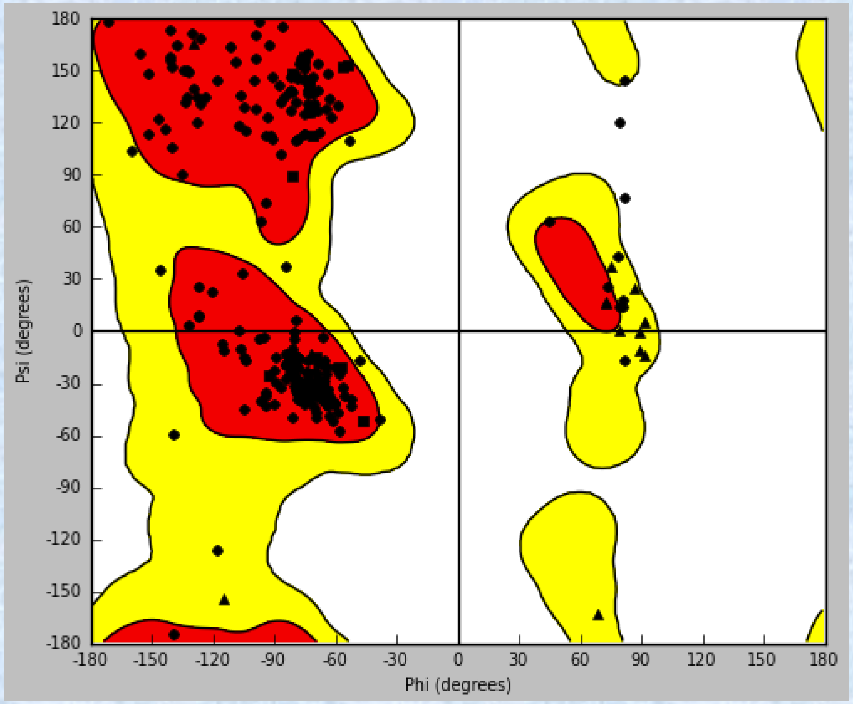


Model validation

Template protein : c-Abl tyrosine kinase
Target protein : c-Yes tyrosine kinase

Identities	Positives	Gaps
197/443 (44%)	281/443 (63%)	15/443 (3%)





		DFG-OUT	DFG-IN
Template PDB id	:	1IEP	1OPK
Resolution	:	2.1 Å	1.80 Å
# of residues in favoured region	:	249 (94.3%)	247 (93.9%)
# of residues in allowed region	:	11 (04.2%)	12 (04.6%)
# of residues in outlier region	:	4 (01.5%)	4 (01.5%)

Web-servers for Homology Modeling

- SWISS Model : http://www.expasy.org/swissmod/SWISS-MODEL.html
- WHAT IF : http://www.cmbi.kun.nl/swift/servers/
- The CPHModels Server : http://www.cbs.dtu.dk/services/CPHmodels/
- 3D Jigsaw : http://www.bmm.icnet.uk/~3djigsaw/
- SDSC1 : http://cl.sdsc.edu/hm.html
- EsyPred3D : http://www.fundp.ac.be/urbm/bioinfo/esypred/

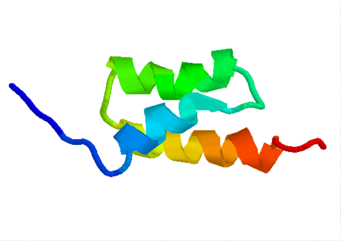
Tools

- COMPOSER http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/matchmaker.html
- MODELER http://salilab.org/modeler
- InsightII http://www.msi.com/
- SYBYL http://www.tripos.com/

Ab-initio Prediction

Prediction from sequence using first principles

AVVTW...GTTWVR ⟶

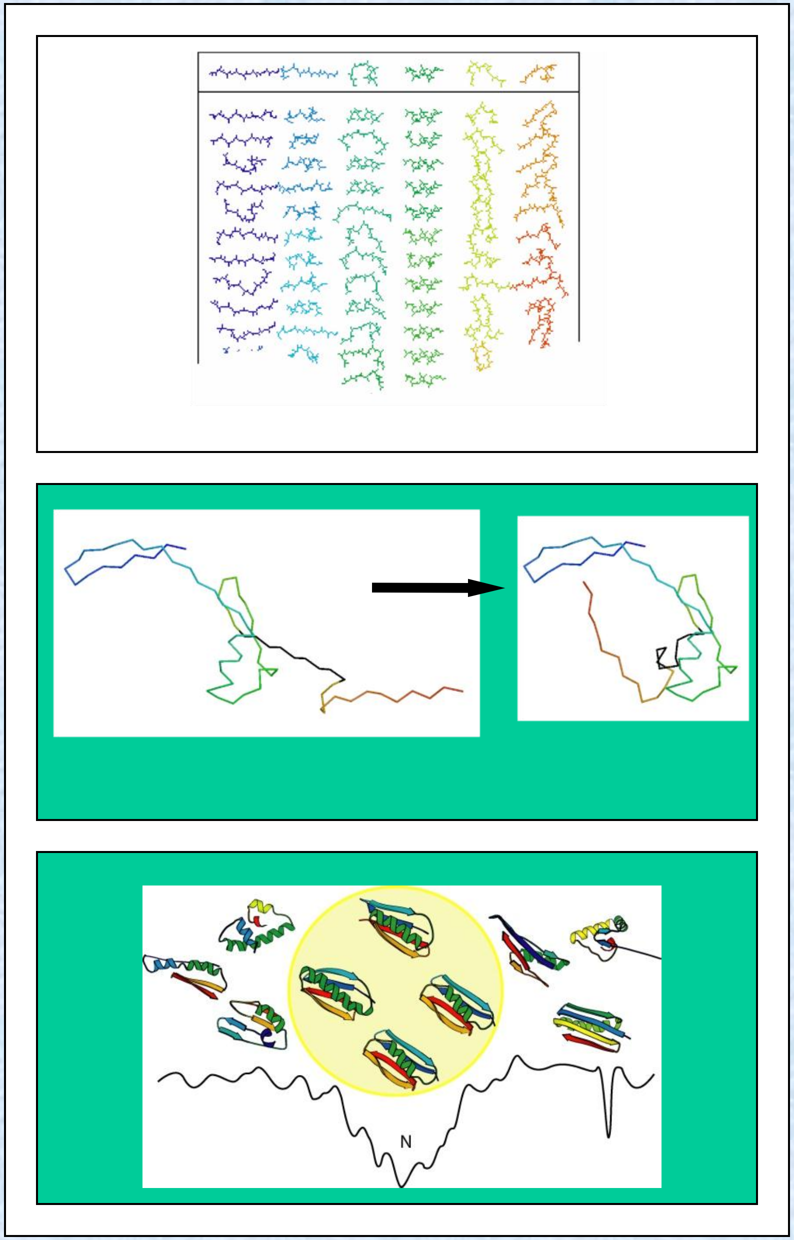


- ***Ab initio*** **protein structure prediction methods build protein 3D structures from sequence based on physical principles.**
- **Importance**
  - **The *ab initio* methods are important even though they are computationally demanding**
  - ***Ab initio* methods predict protein structure based on physical models, they are indispensable complementary methods to Knowledge-based approach**

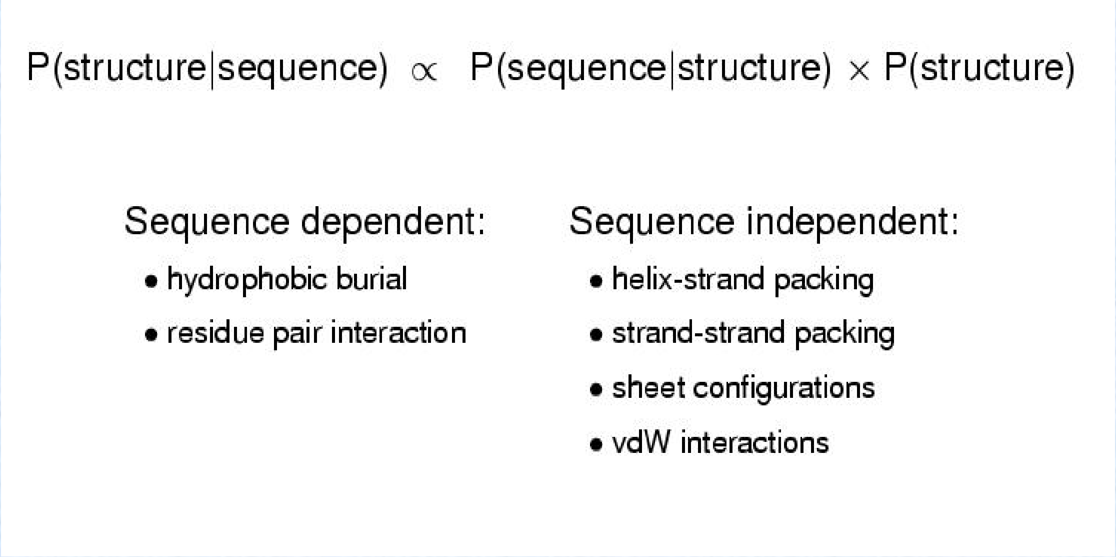
Knowledge-based approach would fail in following conditions:

- **Structure homologues are not available**
- **Possible undiscovered new fold exists**

Structure Prediction with Rosetta



- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys



$$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure})$$

Sequence dependent:

- hydrophobic burial
- residue pair interaction

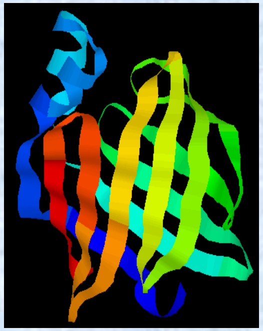
Sequence independent:

- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

Threading - Fold Recognition

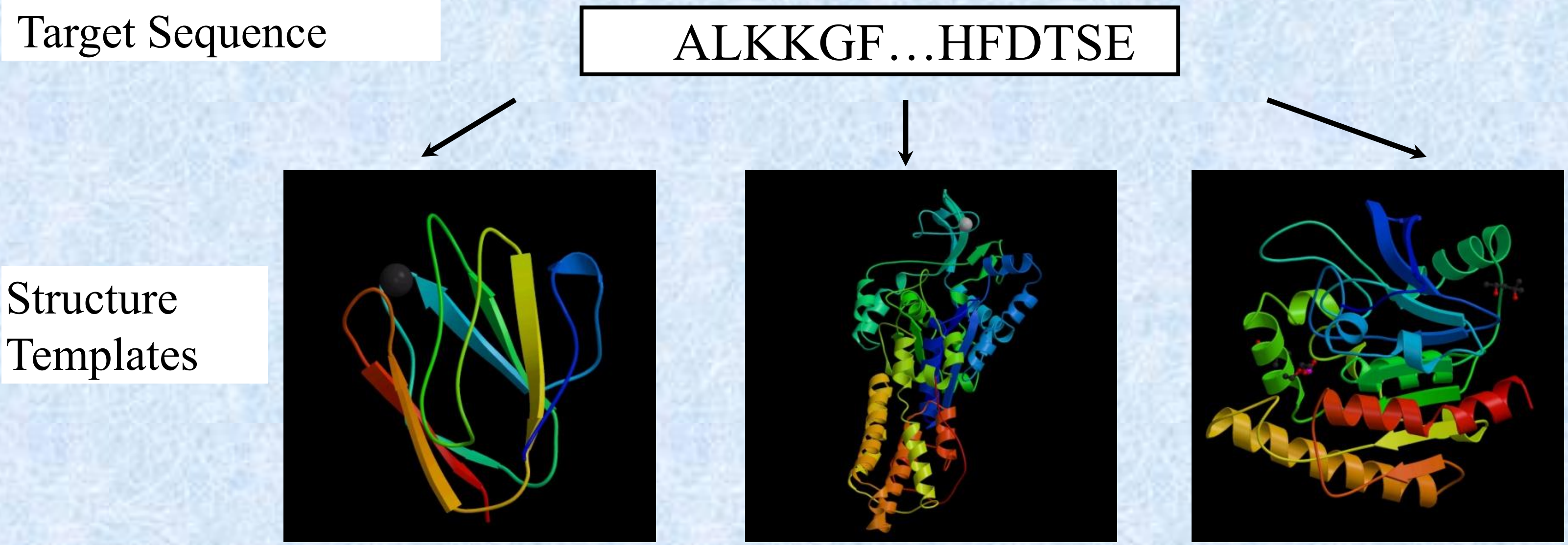
Identify "best" fit between target sequence & template structure

1. **Develop energy function**
2. **Develop template library**
3. **Align target sequence with each template & score**
4. **Identify best scoring template (1D to 3D alignment)**
5. **Refine structure as in homology modeling**



- **Threading - works "sometimes"**
  - **Computationally? Can be expensive or cheap, depends on energy function & whether "all atom" or "backbone only" threading**
  - **Accuracy? in theory, should not depend on sequence identity (should depend on quality of template library & "luck")**
  - **But, usually higher sequence identity ⇒ better model**

Threading



Target Sequence

ALKKGF…HFDTSE

Structure Templates

1. Align target sequence with template structures (fold library) from the Protein Data Bank (PDB)
2. Calculate energy score to evaluate goodness of fit between target sequence & template structure
3. Rank models based on energy scores

Threading Goals & Issues

Find "correct" sequence-structure alignment of a target sequence with its native-like fold in PDB

- *Structure database* - must be complete: no decent model if no good template in library!
- Sequence-structure alignment algorithm:

  Bad alignment ⇒ Bad score!

- Energy function (scoring scheme):
  - must distinguish correct sequence-fold alignment from incorrect sequence-fold alignments
  - must distinguish "correct" fold from close decoys
- Prediction reliability assessment - how determine whether predicted structure is correct (or even close?)

Two main methods (and combinations of these)

Structural profile (environmental)
physicochemical properties of aa's

Contact potential (statistical)

based on contact statistics from PDB

Protein Threading: typical energy function

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

What is "**p**robability" that two specific residues are in contact?

How well does a specific residue fit **s**tructural environment?

Alignment **g**ap penalty?

Total energy: $E_p + E_s + E_g$

Find a sequence-structure alignment that minimizes the energy function

Critical Assessment of Structure Prediction (CASP)

- A Biennial *competition* that has run since 1994.

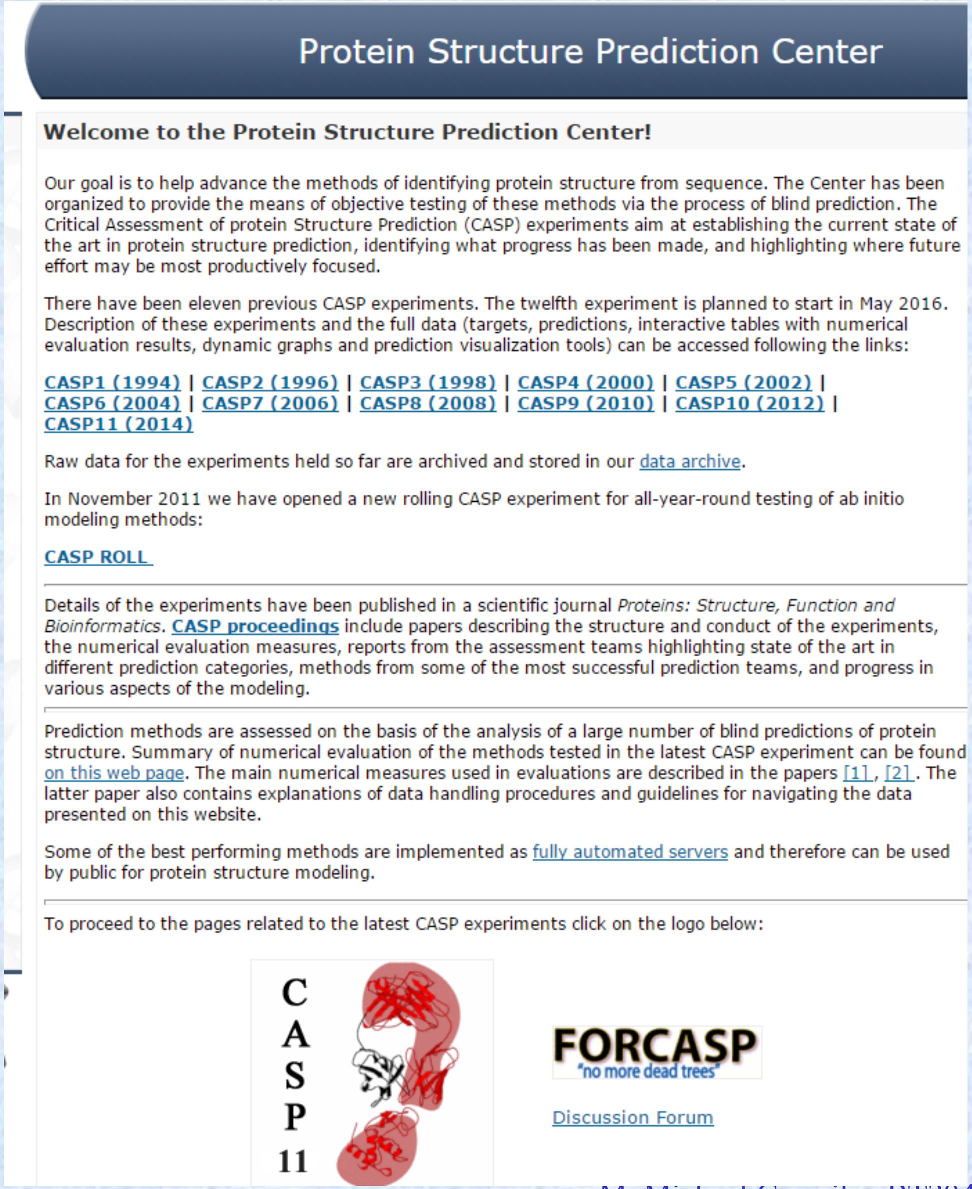
- The next competition will be in 2016 (CASP12)

- http://predictioncenter.org/

- The goal is to advance the methods for predicting protein structure from sequence.

- Protein structures yet to be published are used as blind targets for the prediction methods, with only sequence information released.

- Competitors may use Homology Modelling, Fold recognition or Ab Initio structural prediction methods to propose the structure of the protein.



Protein Structure Prediction Center

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been eleven previous CASP experiments. The twelfth experiment is planned to start in May 2016. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014)

Raw data for the experiments held so far are archived and stored in our data archive.

In November 2011 we have opened a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. CASP proceedings include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1] , [2] . The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

Discussion Forum