Homology modeling with Modeller

Bioinformatics II Presentation
June 20, 2012
Taylor Kessinger
Zarin Shakibaei
Fabio Zanini

Outline

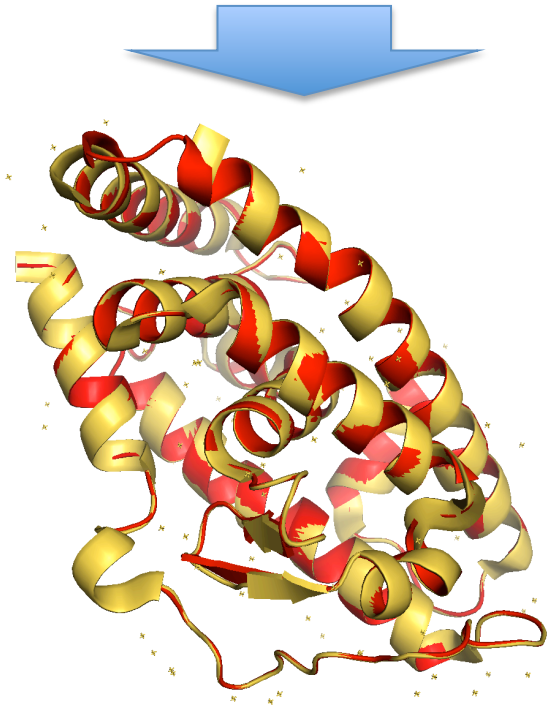
- Goal: Predict the structure of 1yje:A using homology modeling
- Introduction:
  - What is homology modeling?
  - What is Modeller?
- Methods:
  - Initial search
  - Alignment
  - Structure prediction
- Results: Predicted structures and conclusions

Introduction: Homology modeling

- Homologous proteins have similar structures
- Many proteins are homologous to *something* with a known structure
- Structure is *much* more conserved than sequence

```
>P1;1vje_A
structureX:1VJE.pdb: 362 :A:+226 :A:: 2.40: 0.25
----------------------------NLLTSLIRAHLDSGPNTAKLDYSKFQELVLPRFGKEDAGDVQQFYDL
LSGSLDVIRKWAEKIPGFIELSPGDQDLLLESAFLELFILRLAYRSKPGEGKLIFCSGLVLHRLQCARGFGDWID
NILAFSRSLHSLGVDVPAFACLSALVLITDRHGLQDPRRVEELQNRIASCLKEHMAAVS-------CLSRLLGKL
PELRTLCTQGLQRIFCLKLEDLVPPPPIVDKIFMDT---*

>P1;1vje_A
sequence::     : :     : :::-1.00:-1.00
GSSHHHHHHSSGLVPRGSHMQPPDASPTNLLTSLIRAHLDSGPNTAKLDYSKFQELVLPRFGKEDAGDVQQFYDL
LSGSLDVIRKWAEKIPGFIELSPGDQDLLLESAFLELFILRLAYRSKPGEGKLIFCSGLVLHRLQCARGFGDWID
NILAFSRSLHSLGVDVPAFACLSALVLITDRHGLQDPRRVEELQNRIASCLKEHMAAVAGDPQPASCLSRLLGKL
PELRTLCTQGLQRIFCLKLEDLVPPPPIVDKIFMDTLSF*
```



Introduction: Modeller

- Free Python package due to Sali et al. (1994)
- No GUI–knowledge of Python required
- Predicts structures by "satisfaction of spatial restraints"
  - Spatial features common to template structures are globbed
  - Energy or probability functions are assigned to each spatial feature, imposing restraints
  - The model that best satisfies these restrains is selected
  - Examples: molecular PDF, GA341, DOPE (Shen and Sali 2006)

Methods: Initial search

- Target (1yje:A) was CSBlasted against the non-redundant PDB
  - CSBLAST looks for matches by considering the *context* around amino acids (Biegert et al. 2009)
- Hits with $e < 0.01$ were selected as potential templates
  - The first subsequent hit was *much* higher than 0.01, so a clear jump was visible
- An initial MSA was generated using MUSCLE (Edgar 2004)

Methods: Alignment

- Modeller has its own alignment method
  - Required inputs: sequence (in PIR format), .pdb files
  - Gap penalties are lessened outside of secondary structures
- High matches were globbed and placed in a multiple sequence alignment

```
# align the unknown sequence with all templates
seqname = (os.path.basename(seqfile).split('.')[0])
seqname = seqname[:-1]+seqname[-1].upper()
templates = read_templates('../templates/most_relevant.dat')[:n_templates]
templates = filter(lambda t: t['id'] != '1YJE', templates)
n_templates = len(templates)

# check whether the alignment exists already
if not os.path.isfile(seqname+'-multiple_n_'+str(n_templates)+'.ali'):
    aln = alignment(env)
    for template in templates:
        id = template['id']
        chain = template['chain']
        tplname = get_tplname(template)

        mdl = model(env, file=id, model_segment=('FIRST:'+chain, 'LAST:'+chain))
        aln.append_model(mdl, align_codes=tplname, atom_files=id+'.pdb')

    # add the unknown sequence to the alignment.
    aln.append(file=seqfile, align_codes=seqname)

    # Align sequence-structures
    # Note: align2d is obsolete! --> salign
    # Parameters
    # - gap_function makes the gaps in the MSA dependent on the structural
    #   context
    aln.salign(gap_function=True)

    # store the alignment
    aln.write(file=seqname+'-multiple_n_'+str(n_templates)+'.ali', alignment_format='PIR')
```

Methods: Structure prediction

- MSA was used to predict the structure of 1yje:A
  - The automodel class makes this easy
  - More powerful tools available
  - Select best of three models
  - Output: .pdb file

```
# automodel reads the alignment file and actually does the homology modeling for us.
# the output is a .pdb file, to be seen in a.outputs.
# the parameter assess_methods is used to check the quality of the model
# (see also Kryshtafovych and Fidelis, 2009)
a = automodel(env,
              alnfile = seqname+'-multiple_n_'+str(n_templates)+'.ali',
              knowns = map(get_tplname, templates),
              sequence = seqname,
              assess_methods=[getattr(assess, am) for am in assessment_models])

# index of the first/last model
# (determines how many models to calculate)
a.starting_model= 1
a.ending_model = 3

# do the actual homology modeling
a.make()
```

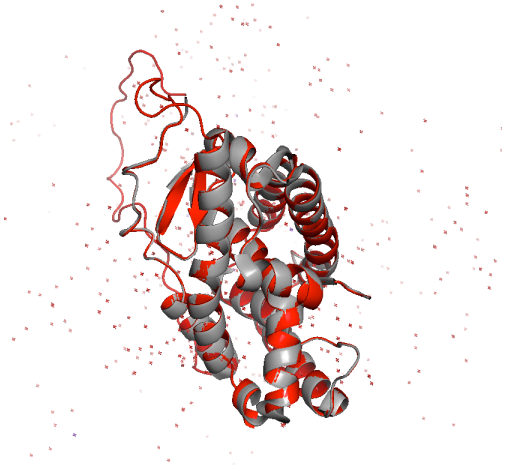
```
EXPDTA    THEORETICAL MODEL, MODELLER 9.10 2012/06/14 15:57:40
REMARK   6 MODELLER OBJECTIVE FUNCTION:      1825.7863
REMARK   6 MODELLER BEST TEMPLATE % SEQ ID:  27.064
ATOM      1  N   GLY     1     -50.586 -28.636 -32.626  1.00 14.91           N
ATOM      2  CA  GLY     1     -51.112 -28.454 -31.254  1.00 14.91           C
ATOM      3  C   GLY     1     -50.015 -28.632 -30.261  1.00 14.91           C
ATOM      4  O   GLY     1     -48.839 -28.659 -30.621  1.00 14.91           O
ATOM      5  N   SER     2     -50.381 -28.748 -28.971  1.00145.17           N
ATOM      6  CA  SER     2     -49.381 -28.916 -27.961  1.00145.17           C
ATOM      7  CB  SER     2     -49.843 -28.497 -26.554  1.00145.17           C
ATOM      8  OG  SER     2     -48.790 -28.684 -25.619  1.00145.17           O
ATOM      9  C   SER     2     -49.009 -30.362 -27.901  1.00145.17           C
```

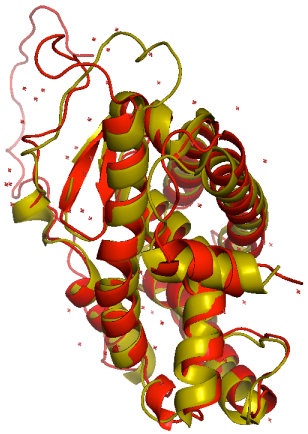
Evaluation of prediction quality

- …can be performed using some quantitative measures
  - Assessments like Krystafovych et al. (2009)
  - molpdf, GA341 and DOPE (all built-in)
- …can be done visually in PyMOL
  - With true structure (Flaig et al. 2005)
  - Without

Results: Prediction checks

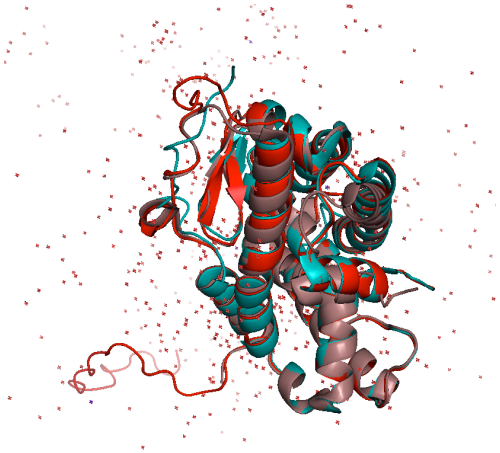
- Top: 1yje:A (red) fitted onto one template (1ovl:A)
- Bottom: Predicted structure (red) compared to real one
- Note poor fitting of loops, but overall good prediction of secondary structures
- DOPE score: -31212





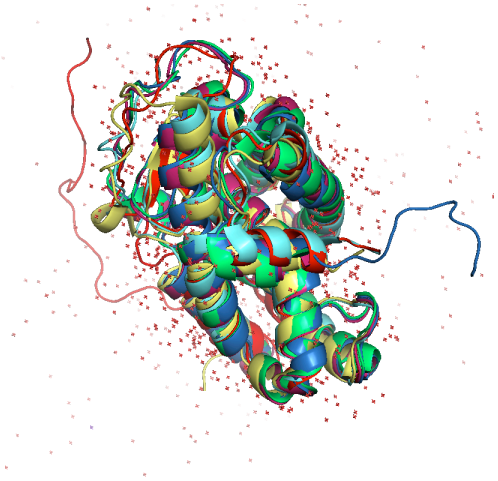
Results: Prediction checks

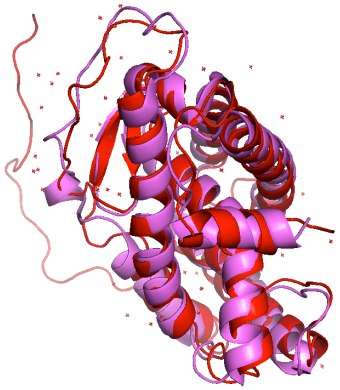
- Top: 1yje:A (red) fitted onto two templates (1ovl:A and 1pdu:A)
- Bottom: Predicted structure (red) compared to real one
- Loop fitting appears to have improved slightly
- DOPE score: -31385



Results: Prediction checks

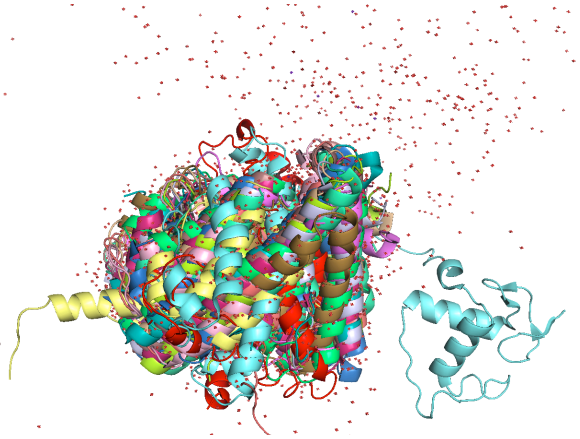
- Top: 1yje:A (red) fitted onto five templates
- Bottom: Predicted structure (red) compared to real one
- Loop fitting is now pretty good, overall
- DOPE score: -30780

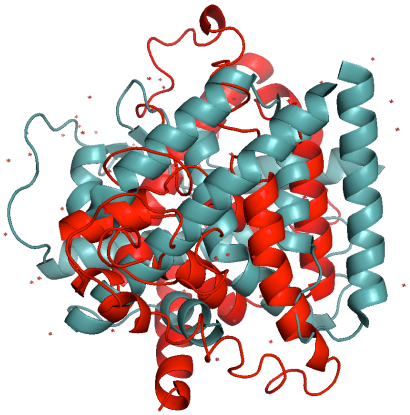




Results: Prediction checks

- Top: 1yje:A (red) fitted onto fifteen templates
- Bottom: Predicted structure (red) compared to real one
- There appear to be diminishing returns
- DOPE score: -21209
- Notice the jump!





Conclusions

- Modeller can accurate predict secondary structures and even loop regions
  - Performs well when it doesn't get confused
  - Some level of skill is needed in selecting matches
- Current PDB may be enough to solve structure prediction problem (Zhang 2005)
  - Homology modeling may be the way of the future; alignment and selection are limiting
  - Just a bit too ambitious

Acknowledgments

- Dr. Sven Nahnsen
- Remainder of the Bio II class

References

- Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.
- A. Biegert and J. Söding. Sequence context-specific profiles for homology searching. *Proceedings of the National Academy of Sciences*, 106(10):3770– 3775, 2009.
- R.C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.
- Ralf Flaig, Holger Greschik, Carole Peluso-Iltis, and Dino Moras. Structural basis for the cell-specific activities of the ngfi-b and the nurr1 ligand- binding domain. *Journal of Biological Chemistry*, 280(19):19250–19258, 2005.
- Andriy Kryshtafovych and Krzysztof Fidelis. Protein structure prediction and model quality assessment. *Drug Discovery Today*, 14(78):386 –393, 2009.
- A. Sali and TL Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Protein Structure by Distance Analysis*, 64:C86, 1994.
- Andrej Sali, Liz Potterton, Feng Yuan, Herman van Vlijmen, and Martin Karplus. Evaluation of comparative protein modeling by modeller. *Proteins: Structure, Function, and Bioinformatics*, 23(3):318–326, 1995.
- Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.2r3pre.
- Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, 2006.
- Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current pdb library. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1029, 2005.