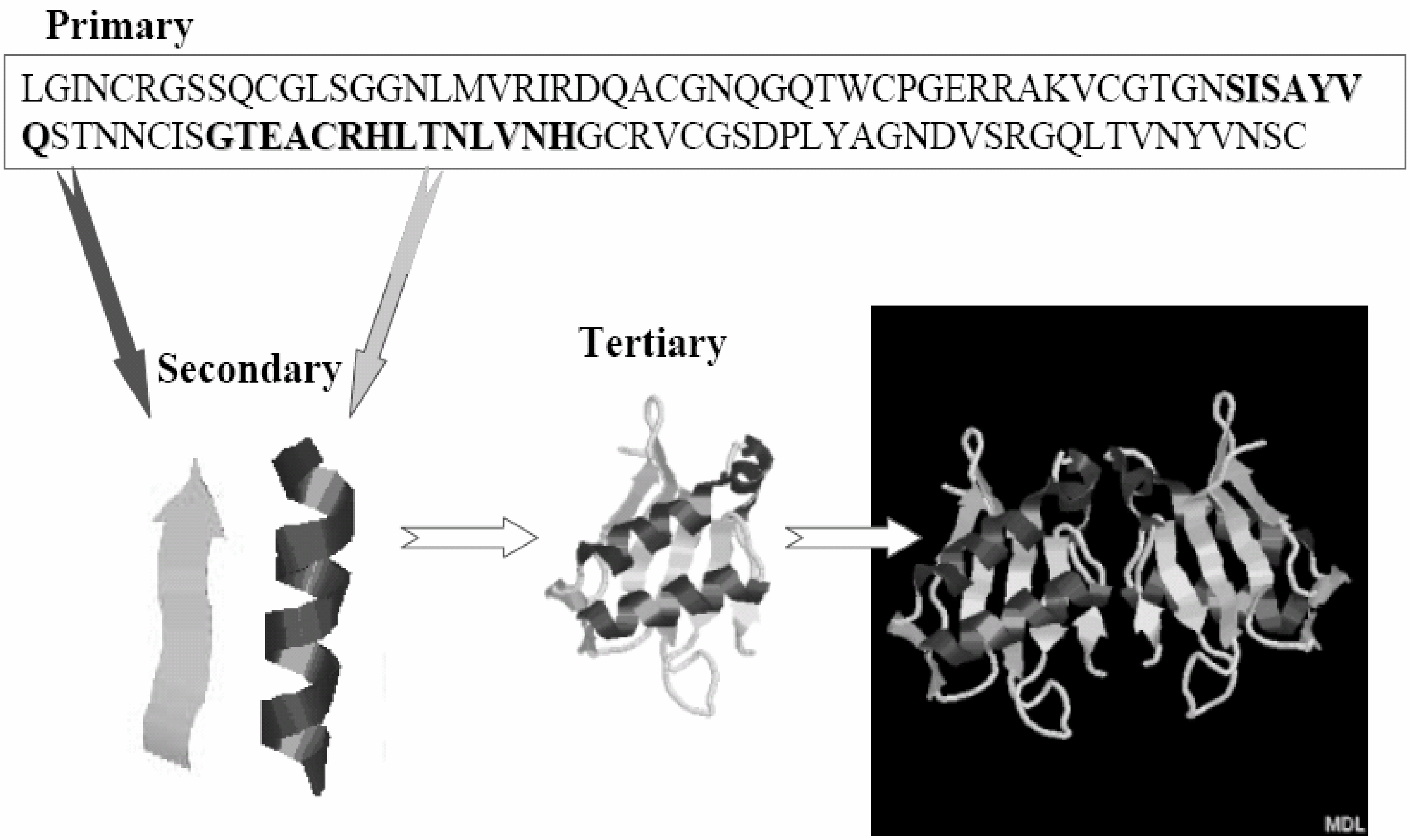
Homology Modeling and Structural analysis

-- some basic concepts and examples

Ye Zhiqiang

Levels of Protein Structure



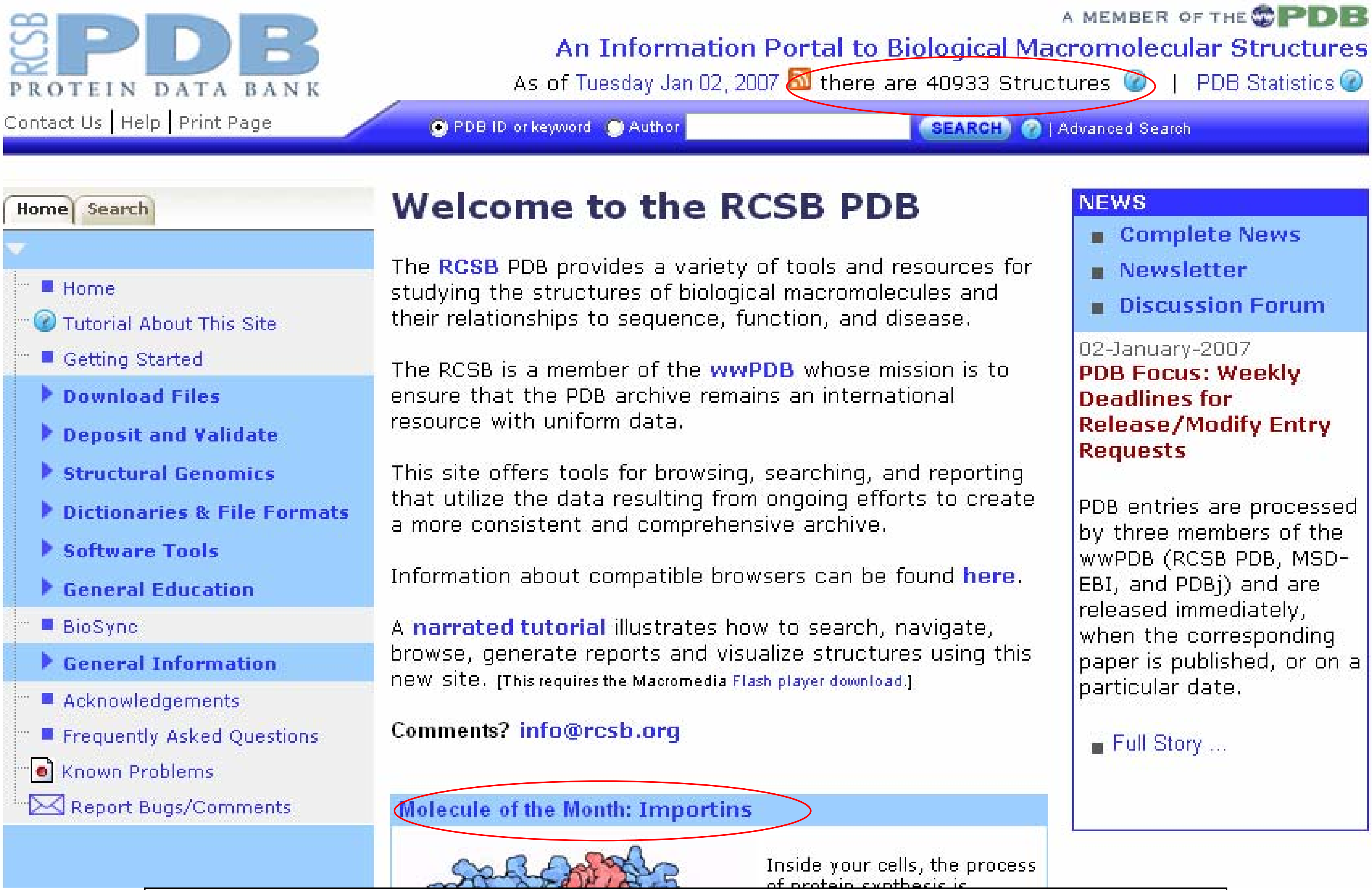
Primary

LGINCRGSSQCGLSGGNLMVRIRDQACGNQGQTWCPGERRAKVCGTGN**SISAYV**
QSTNNCIS**GTEACRHLTNLVNH**GCRVCGSDPLYAGNDVSRGQLTVNYVNSC

Secondary

Tertiary

Protein Data Bank



An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 02, 2007 there are 40933 Structures | PDB Statistics

Contact Us | Help | Print Page

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Importins

Inside your cells, the process of protein synthesis is

NEWS

- Complete News
- Newsletter
- Discussion Forum

02-January-2007

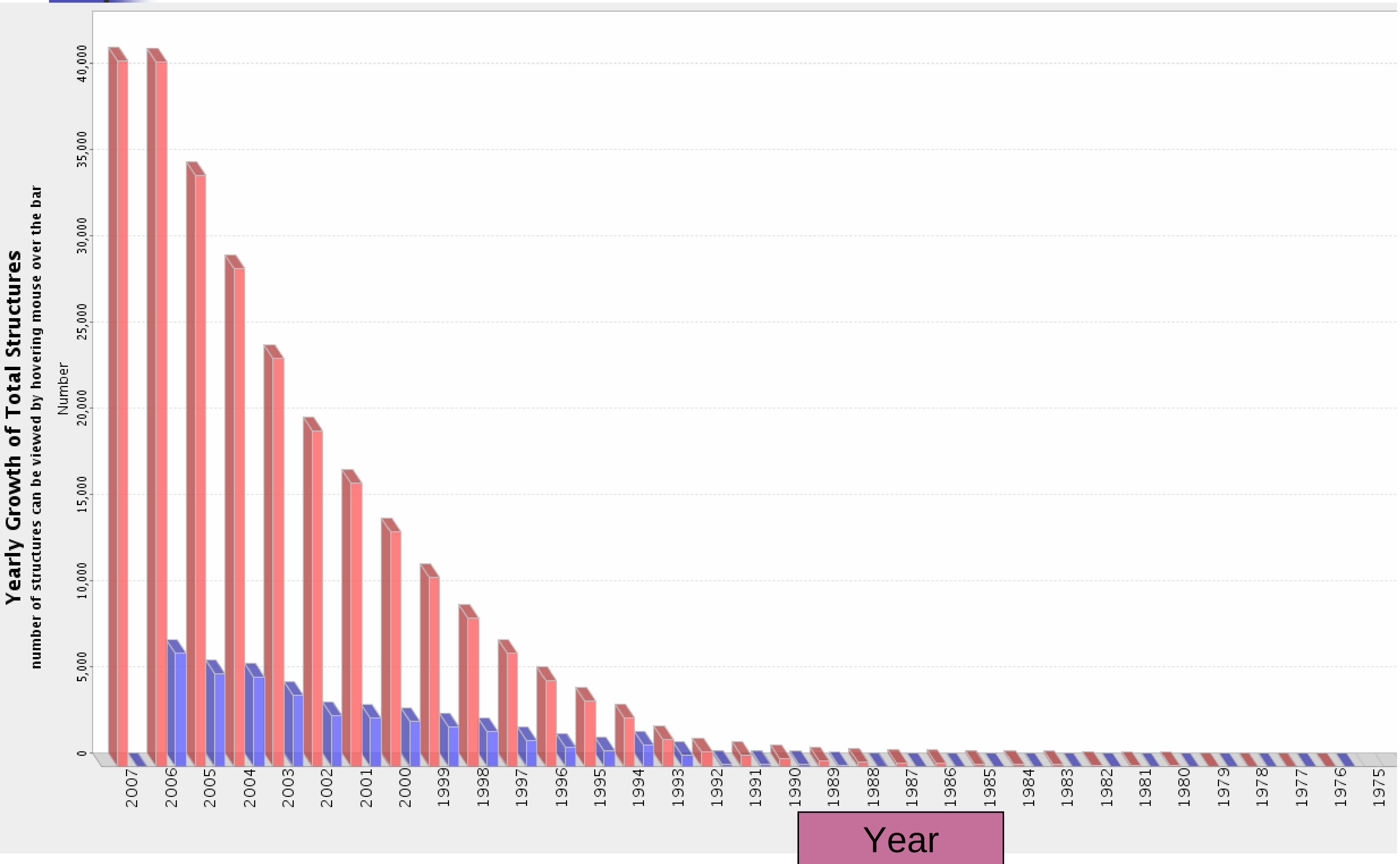
PDB Focus: Weekly Deadlines for Release/Modify Entry Requests

PDB entries are processed by three members of the wwPDB (RCSB PDB, MSD-EBI, and PDBj) and are released immediately, when the corresponding paper is published, or on a particular date.

- Full Story ...

http://www.rcsb.org/pdb

The Growth of PDB entries

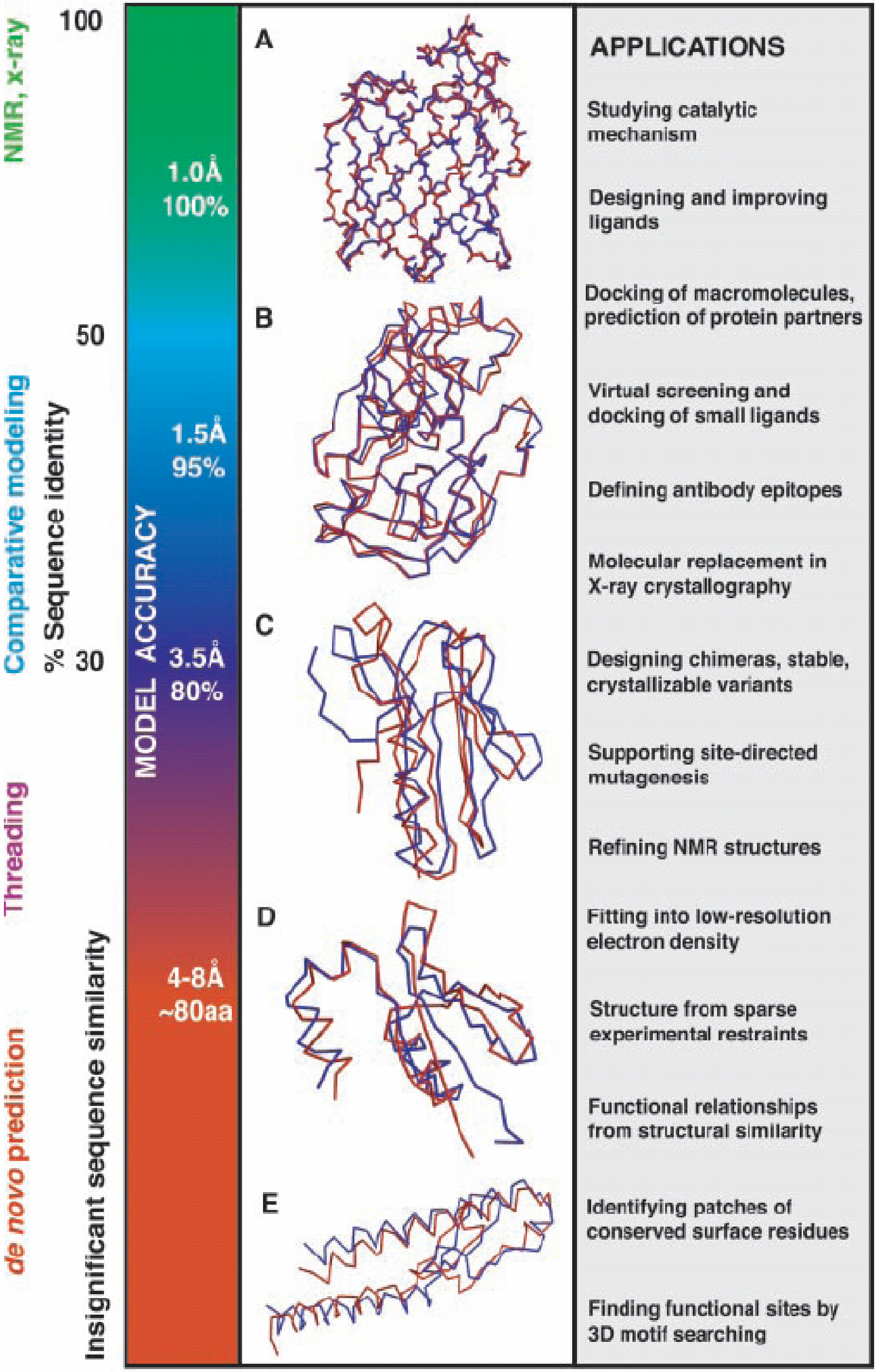


But …

- The growth of protein structures falls behind that of protein sequence largely.
- The new sequenced genes require a proper prediction of the structure of its protein product, for the functional prediction.

What is homology modeling

- Build the structure of the structure-unknown protein (target) according to a proper protein with known-structure (template).
- The template and target should have a considerable sequence identity.
- This is based on the rule: sequence determines the structure.


NMR, x-ray
Comparative modeling
Threading
de novo prediction
% Sequence identity
Insignificant sequence similarity
100
50
30
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching

- Target 和template之间序列同一性(identity)高低不同，相应的结构预测方法也随之变化.
- 不同的序列同一性条件下，预测出来的结构模型的可应用范围也不同.
- 自动同源建模一般要求序列同一性超过35%，这样结果比较可靠。

From Science 2001,
Baker & Sali

Why we require a sufficient sequence identity?

- The quality of the model is mainly determined by the sequence alignment between target and template.
- The automatic sequence alignment methods often fail to generate good enough alignments because of the low sequence identity.
- So ...

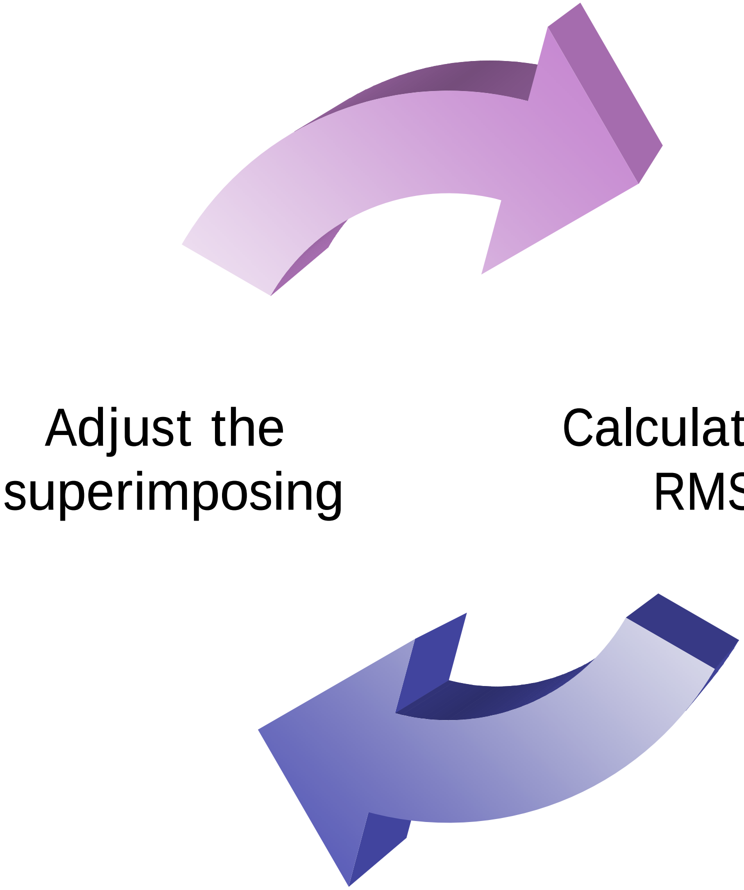
•However, combining other evidences to improve the alignment manually will generate rather good models, even if the identity is low.

Structure comparison and similarity

- Superimpose (structural alignment)
  - Translation (平动)
  - Rotation (转动)
- Root Mean Square Deviation (RMSD)
  - In a structure **alignment** RMSD measures how far the **aligned atoms** are from each other on average

Structural equivalent atoms, or say, counterparts

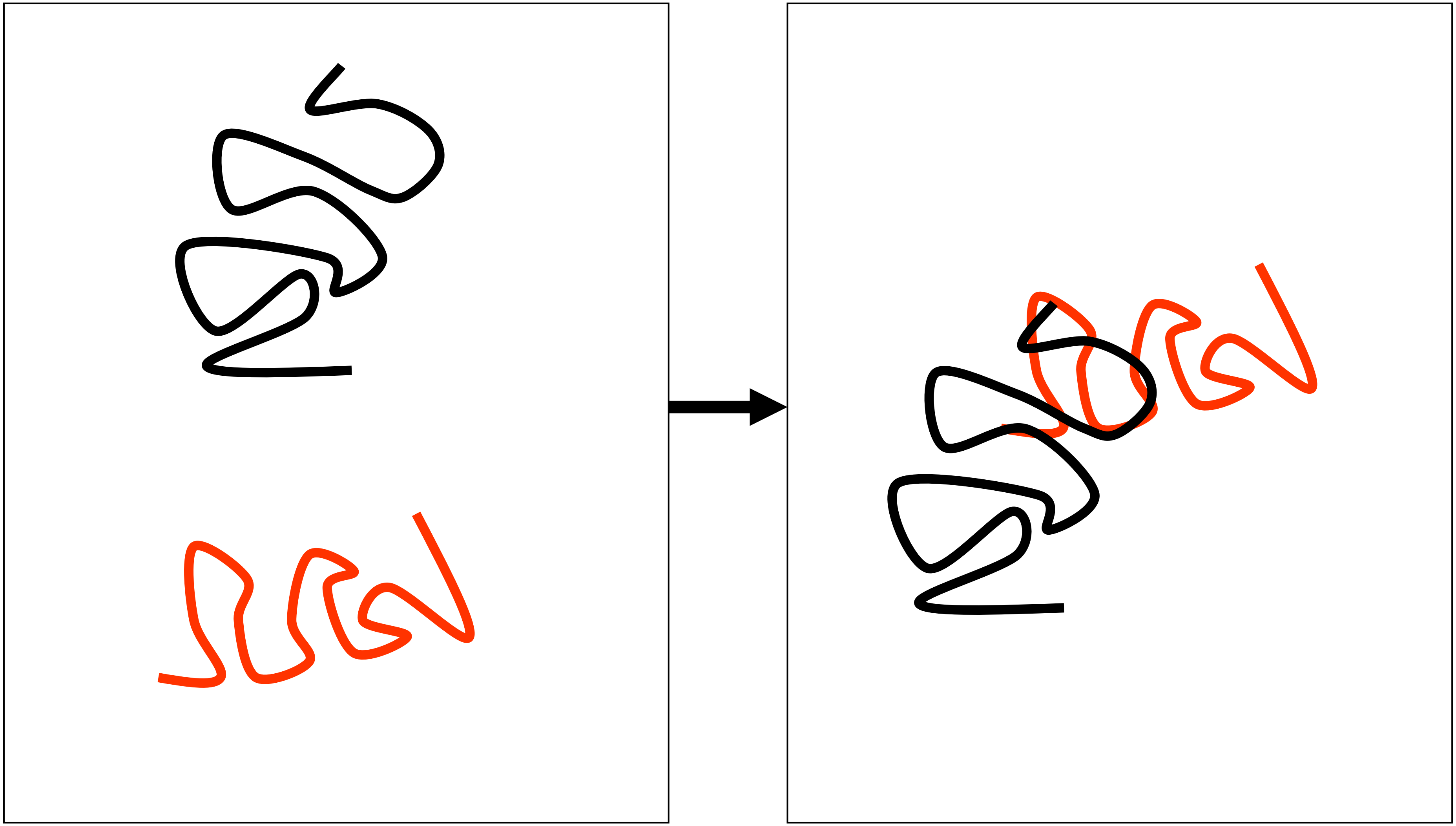
Iterate until the convergence



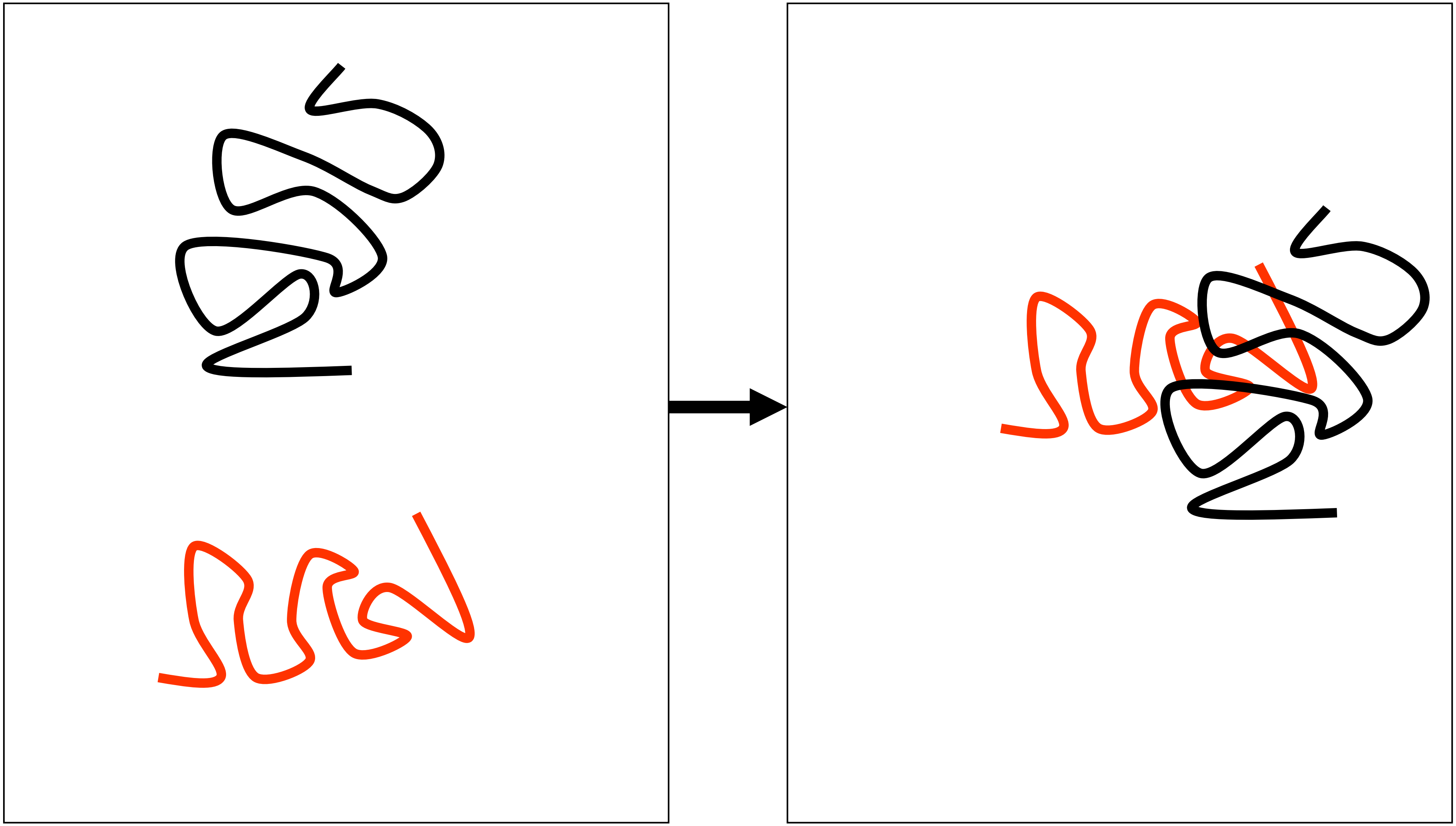
Adjust the superimposing

Calculate the RMSD

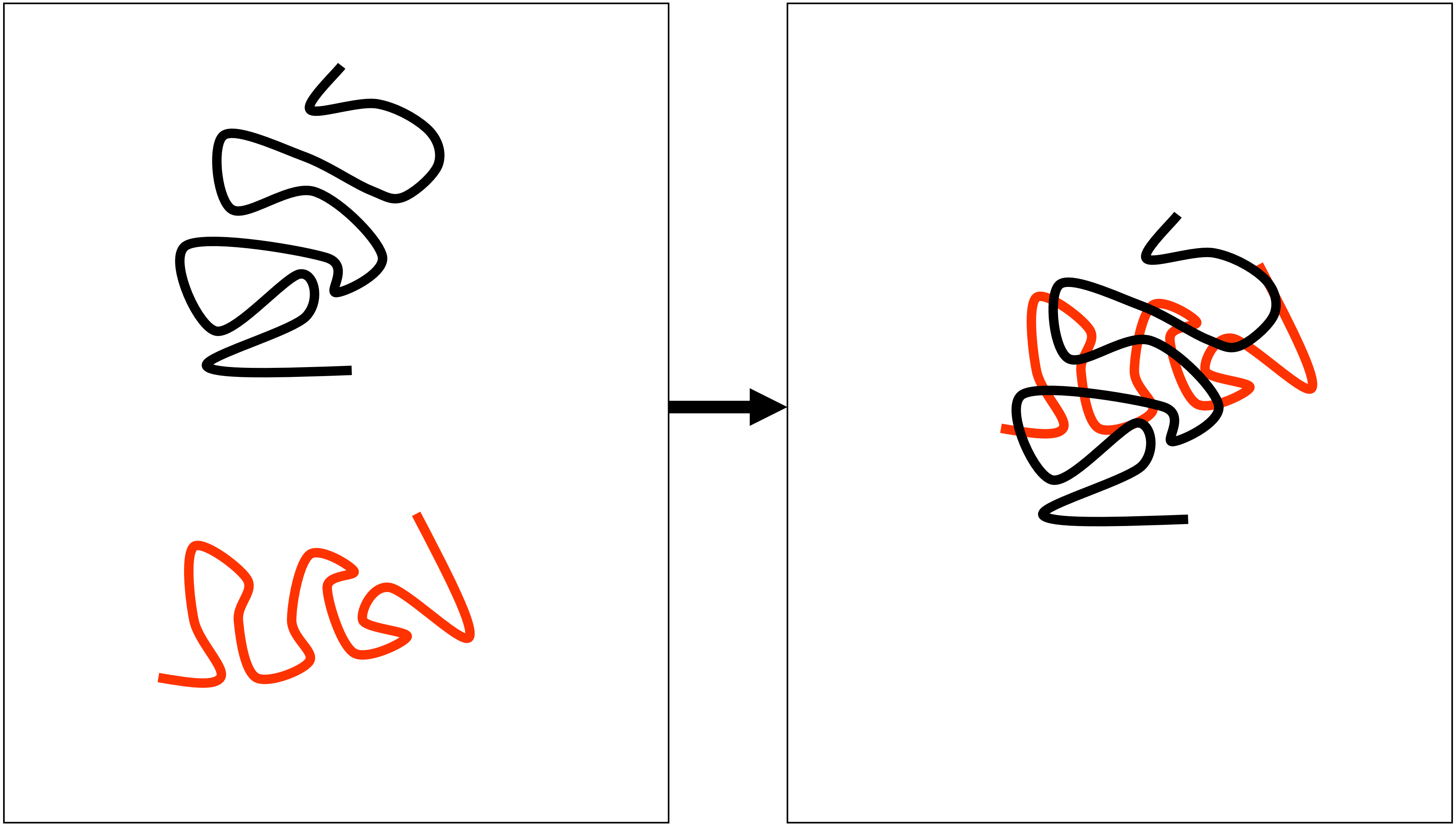
Basic Operations: Translation



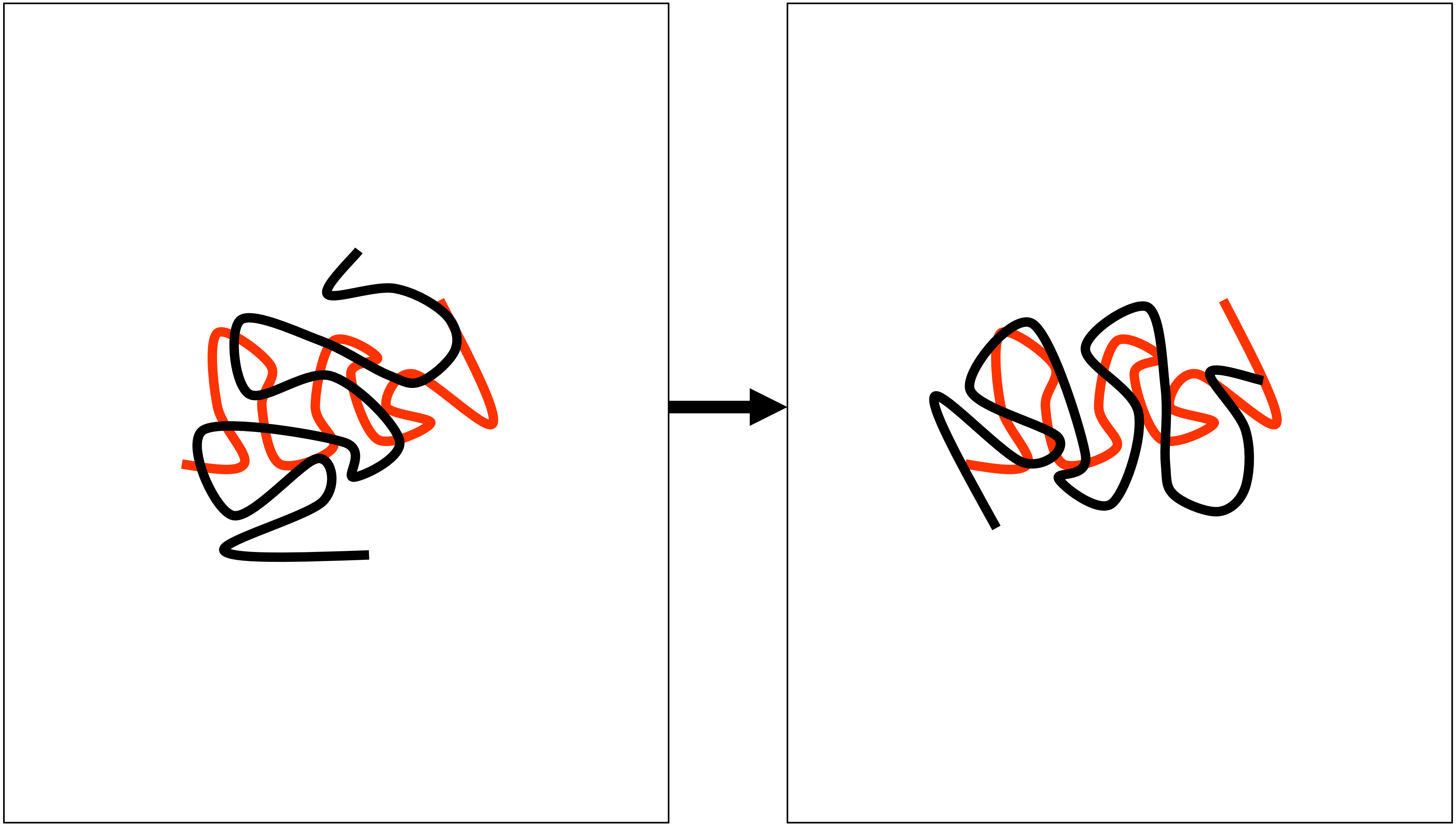
Basic Operations: Translation



Basic Operations: Translation



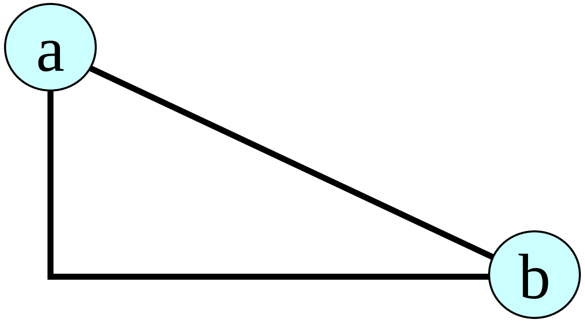
Basic Operations: Rotation



Root Mean Square Deviation

- What is the distance between two points *a* $(x_a, y_a)$and *b* (x_b, y_b)
  - Euclidean distance:

$$d(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$



  - In 3D space:

$$d(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$$

Root Mean Square Deviation

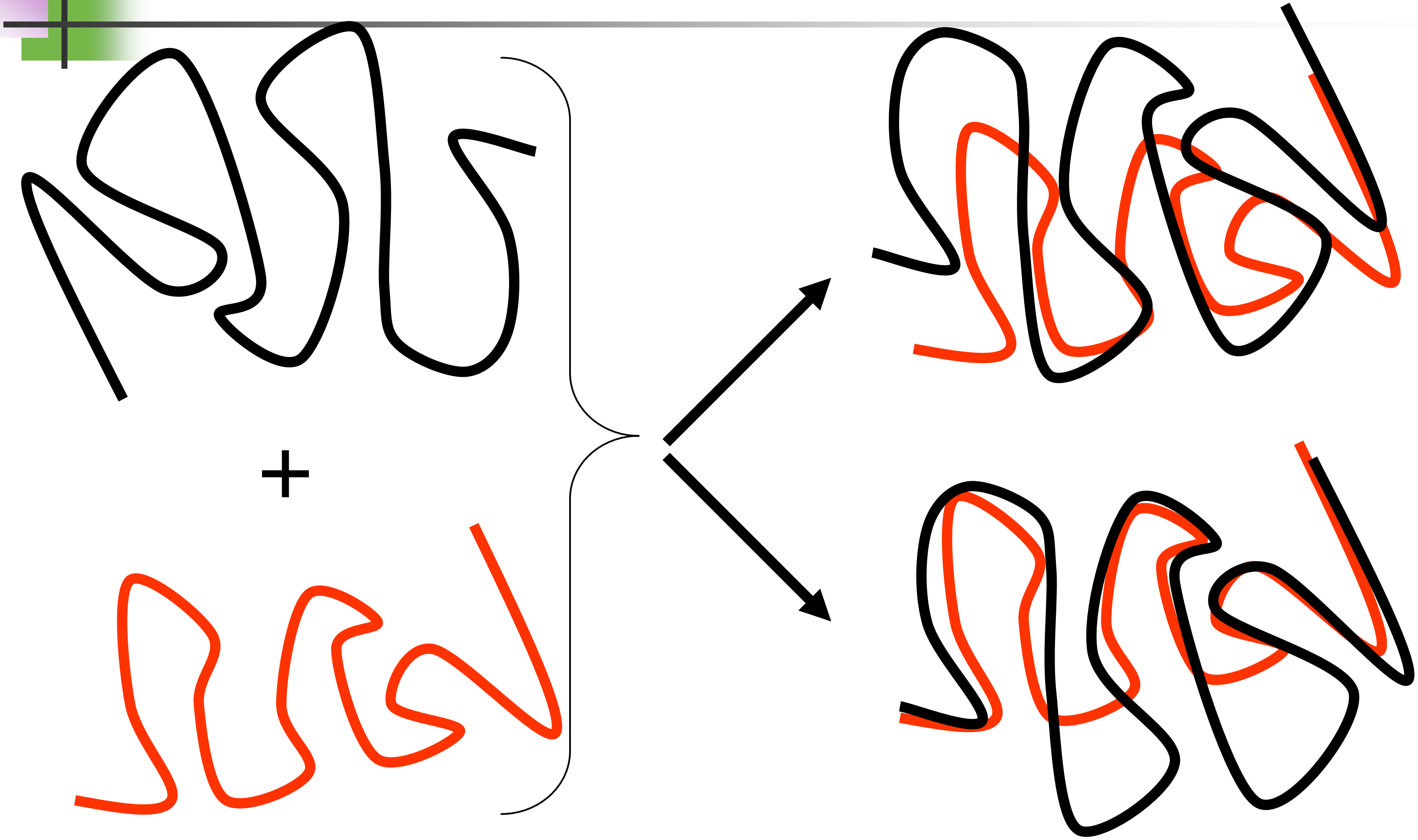
- After the structural alignment (superimpose), *d_i* represents the distance between the i*th* aligned atom pair (there are n pairs in total), the root mean square deviation is defined as:

$$rmsd = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

Quality of Alignment

- Identical structures => *RMSD* = “0”
- Similar structures => *RMSD* is small (1 – 3 Å)
- Distant structures => *RMSD* > 3 Å

Structure Alignment (open problem)



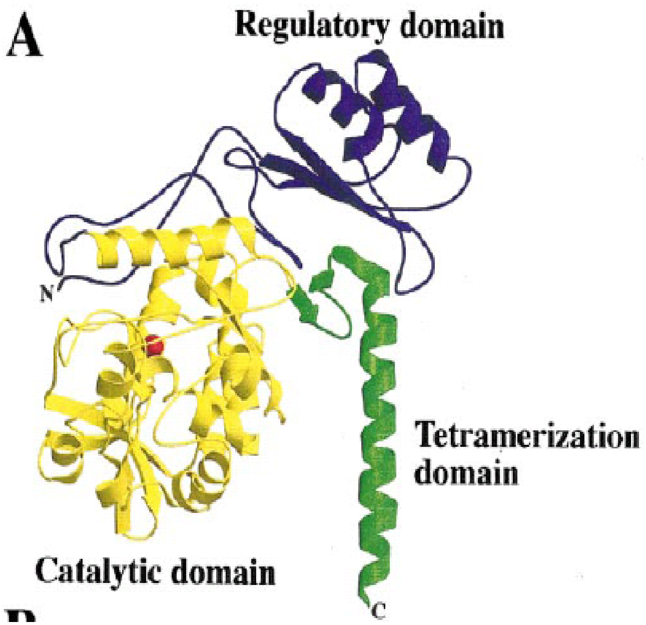
Examples of homology modeling

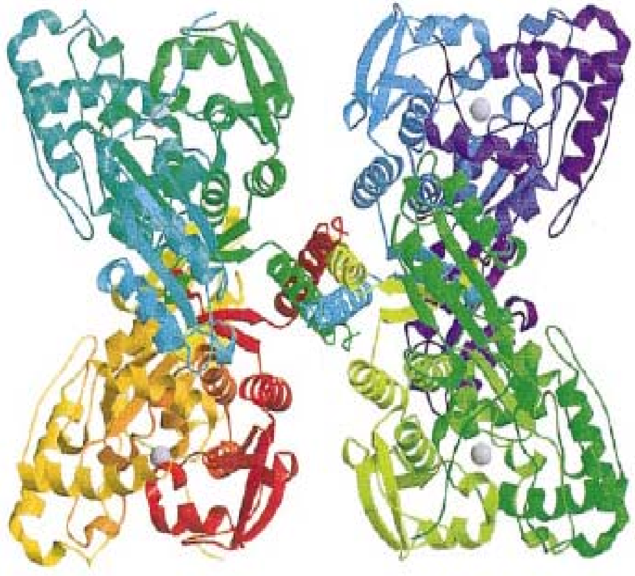
- Human PKU -- structural assembly
- Rice EPSPS -- comparative modeling

Background of Phenylalanine Hydroxylase (PAH)

- Locus: 12q24.1
- EC: 1.14.16.1
- Catalysis: Phe → Tyr
- Phe是人体必需氨基酸，人自身不能合成，需要从食物中获得。但食物中Phe往往过量，所以需要PAH来催化其转为Tyr.
- 当PAH工作不正常的时候，血液中Phe浓度增高，影响幼儿智力发育。Phe进入其代谢旁路，形成苯丙酮酸（phenyl ketonuria)从尿液排出，简称PKU症或者HPA症）
- 人群中发病率1/10000，autosomal recessive

PAH protein



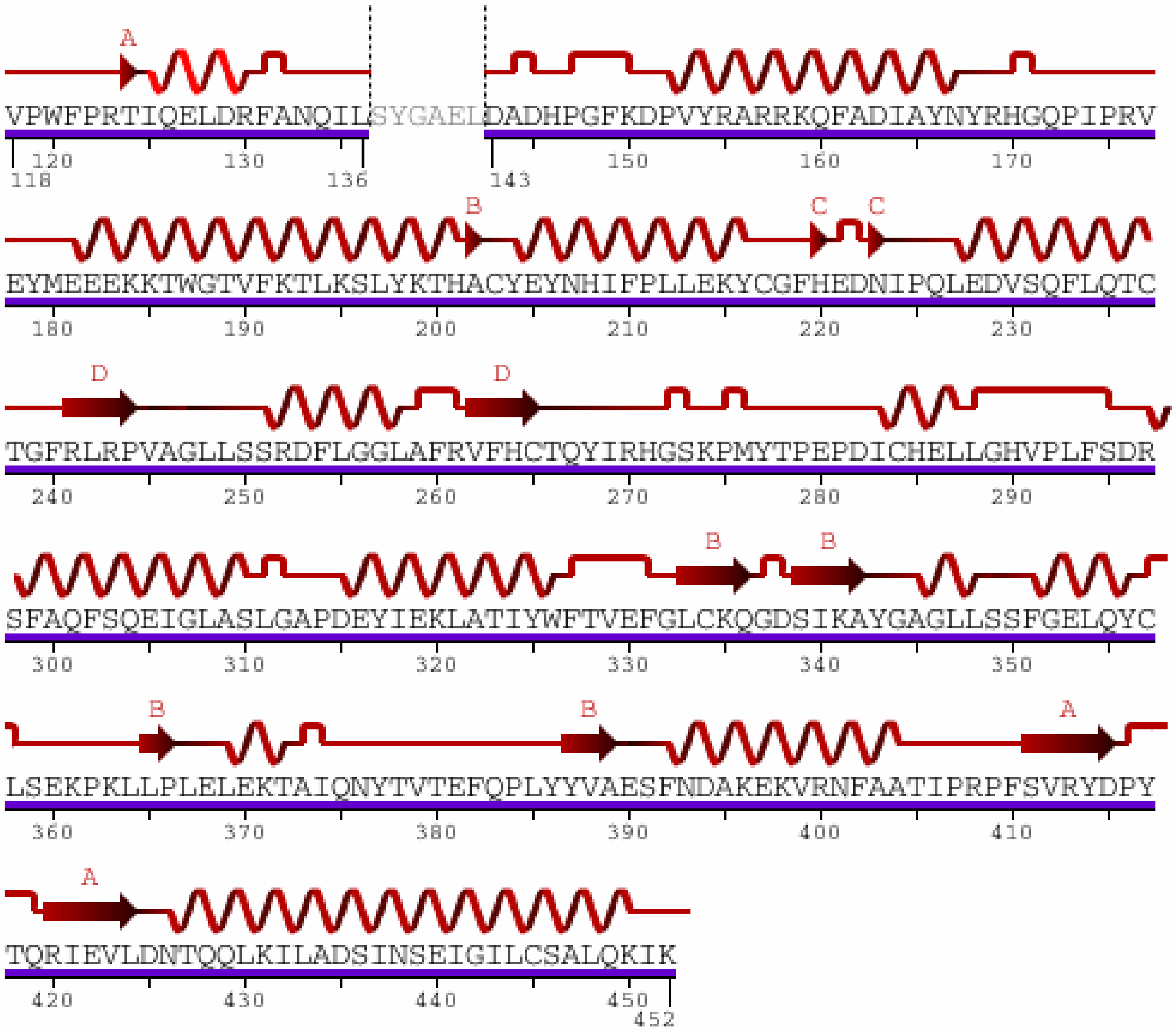


- 在蛋白水平上，总共长452aa，从N端到C端有3个domain：调节域，催化域和四聚体域。
- 其有功能的形式是四聚体

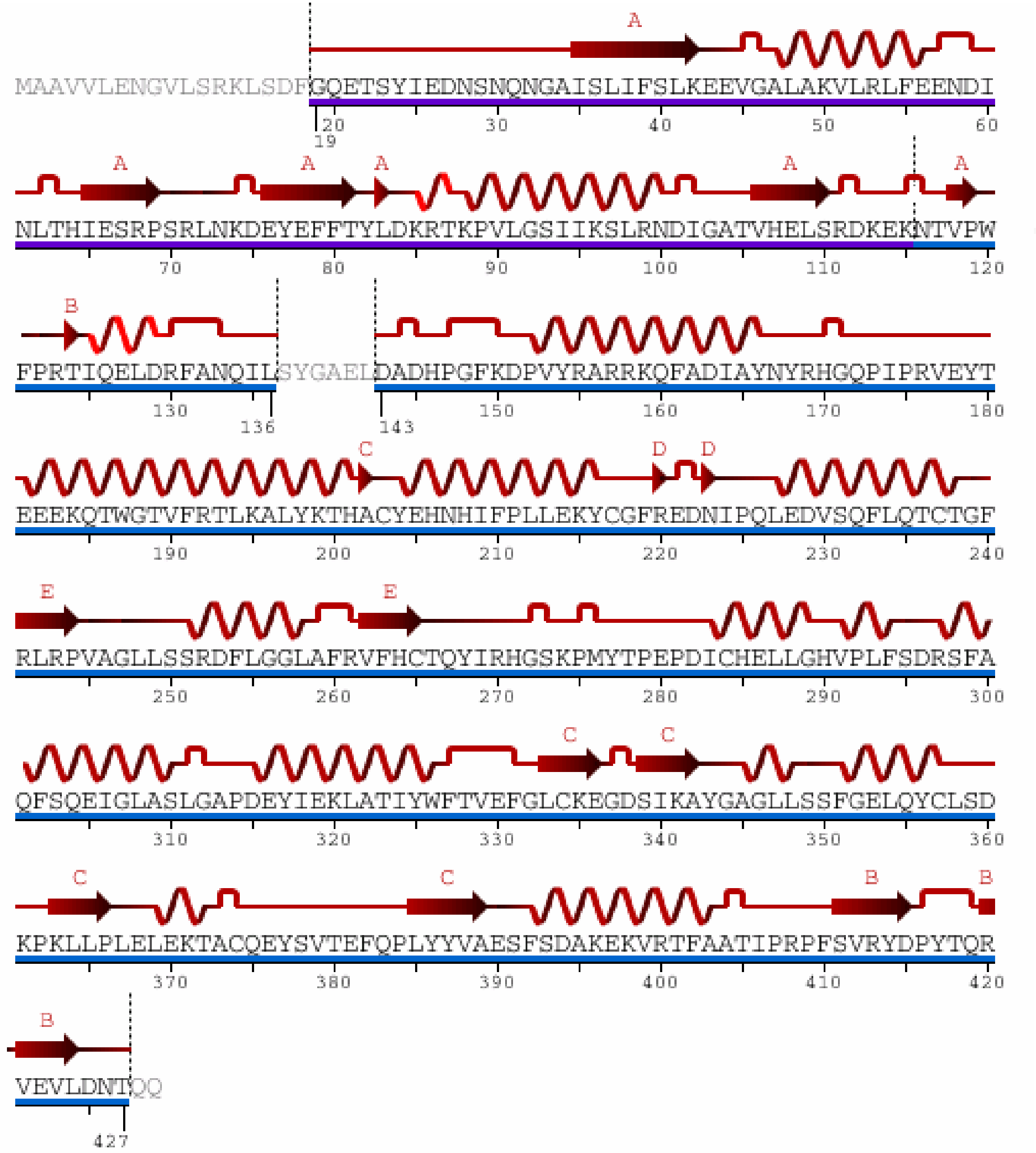
现有的两个PAH的结构

- 2PAH： tetramer, 118-452, 其中仍然有少量残基的坐标缺失，比如两条链的137-142, 另外两条链的131-143 (disordered region?)
- 1PHZ: dimer, 1-429，但是其中的1-18, 137-142, 428-429缺失

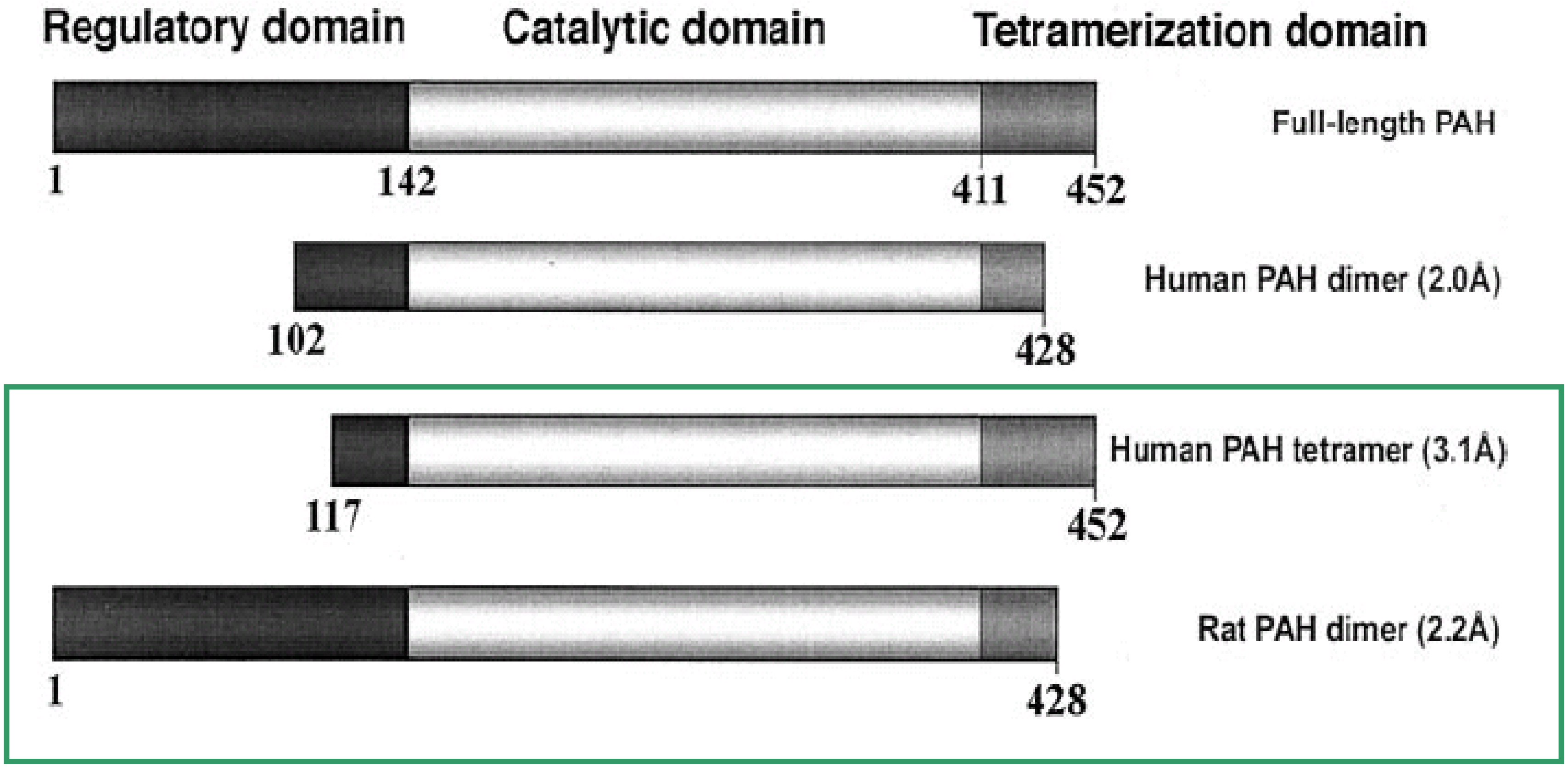
2PAH


A
VPWFPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRV
118 120 130 136 143 150 160 170
B C C
EYMEEEKKTWGTVFKTLKSLYKTHACYEYNHIFPLLEKYCGFHEDNIPQLEDVSQFLQTC
180 190 200 210 220 230
D D
TGFRLRPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDR
240 250 260 270 280 290
B B
SFAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKQGDSIKAYGAGLLSSFGELQYC
300 310 320 330 340 350
B B A
LSEKPKLLPLELEKTAIQNYTVTEFQPLYYVAESFNDAKEKVRNFAATIPRPFSVRYDPY
360 370 380 390 400 410
A
TQRIEVLDNTQQLKILADSINSEIGILCSALQKIK
420 430 440 450 452

1PHZ


A
MAAVVLENGVLSRKLSDFGQETSYIEDNSNQNGAISLIFSLKEEVGALAKVLRLFEENDI
20 30 40 50 60
19
A A A A A
NLTHIESRPSRLNKDEYEFFTYLDKRTKPVLGSIIKSLRNDIGATVHELSRDKEKNTVPW
70 80 90 100 110 120
B
FPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRVEYT
130 136 143 150 160 170 180
C D D
EEEKQTWGTVFRTLKALYKTHACYEHNHIFPLLEKYCGFREDNIPQLEDVSQFLQTCTGF
190 200 210 220 230 240
E E
RLRPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDRSFA
250 260 270 280 290 300
C C
QFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKEGDSIKAYGAGLLSSFGELQYCLSD
310 320 330 340 350 360
C C B B
KPKLLPLELEKTACQEYSVTEFQPLYYVAESFSDAKEKVRTFAATIPRPFSVRYDPYTQR
370 380 390 400 410 420
B
VEVLDNTQQ
427

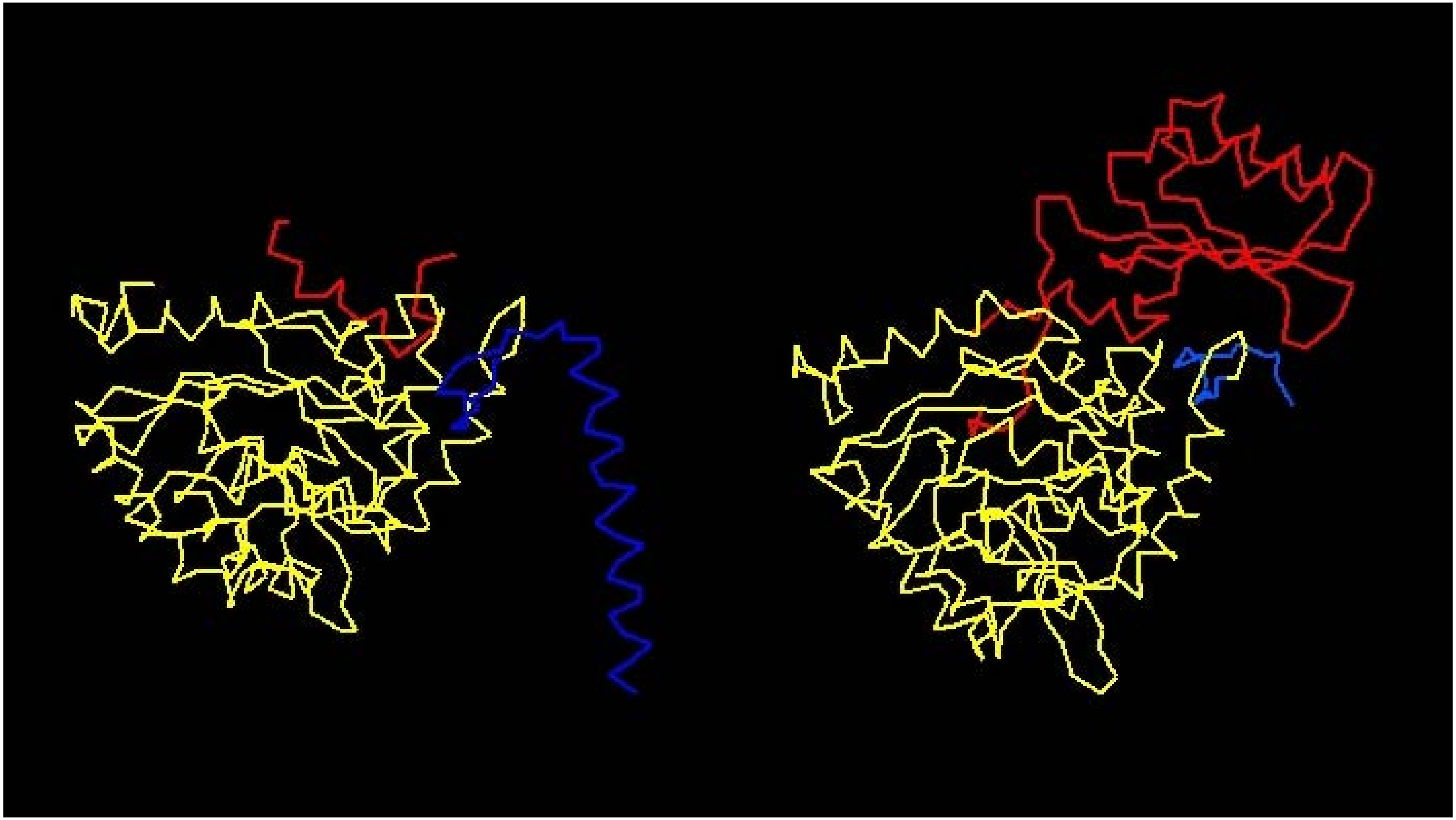
Alignment


Regulatory domain
Catalytic domain
Tetramerization domain
Full-length PAH
1
142
411
452
Human PAH dimer (2.0Å)
102
428
Human PAH tetramer (3.1Å)
117
452
Rat PAH dimer (2.2Å)
1
428

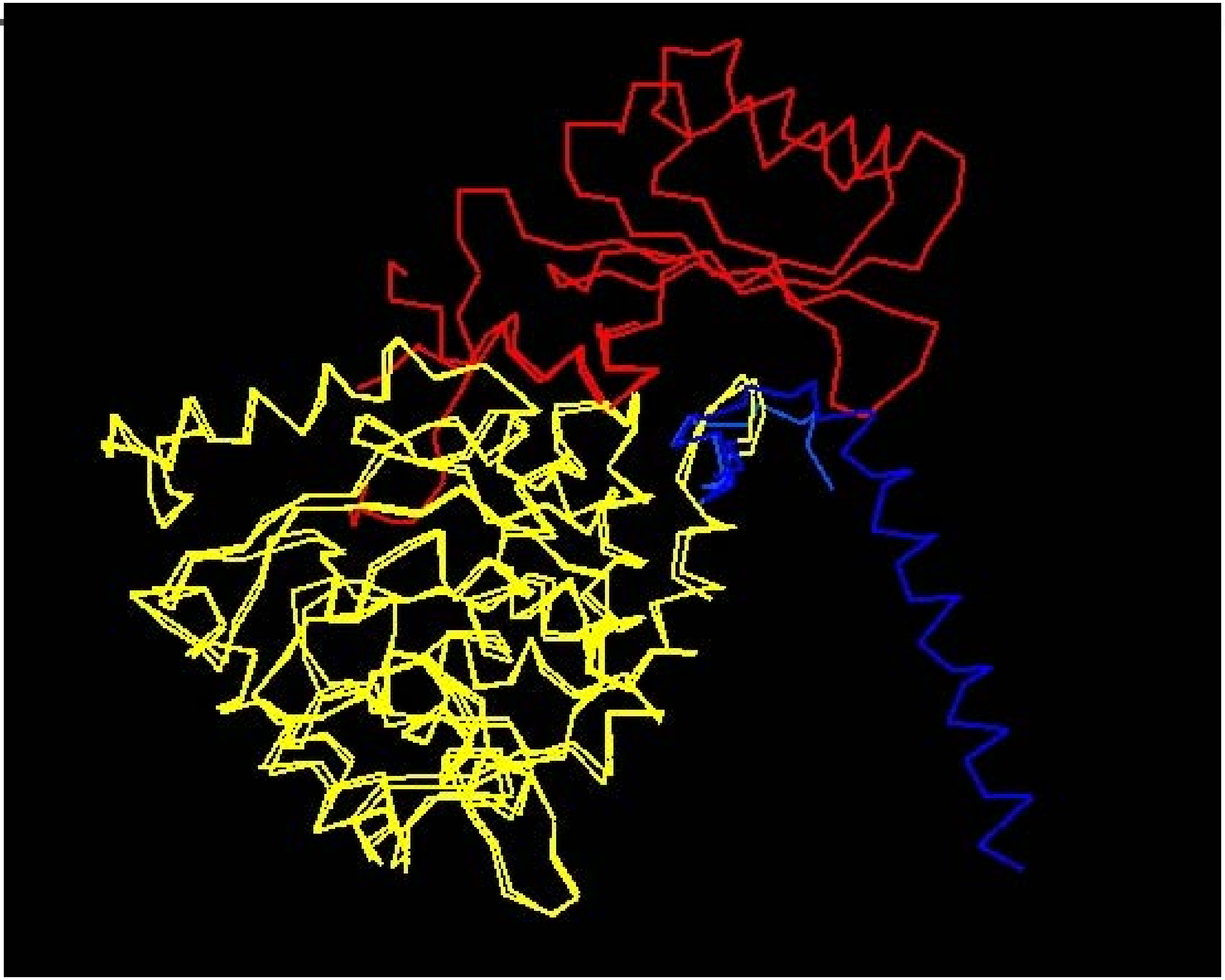
Some points:

- 用2PAH的四聚体形式作为模板是必要的，这样在其每一个单体上面都可以和1PHZ的一个单体进行重叠部分的叠合，这样最后可以得到一个接近全长的四聚体的模板结构。
- 下面的图示都只用了单体，是为了更清晰的展示这个过程。

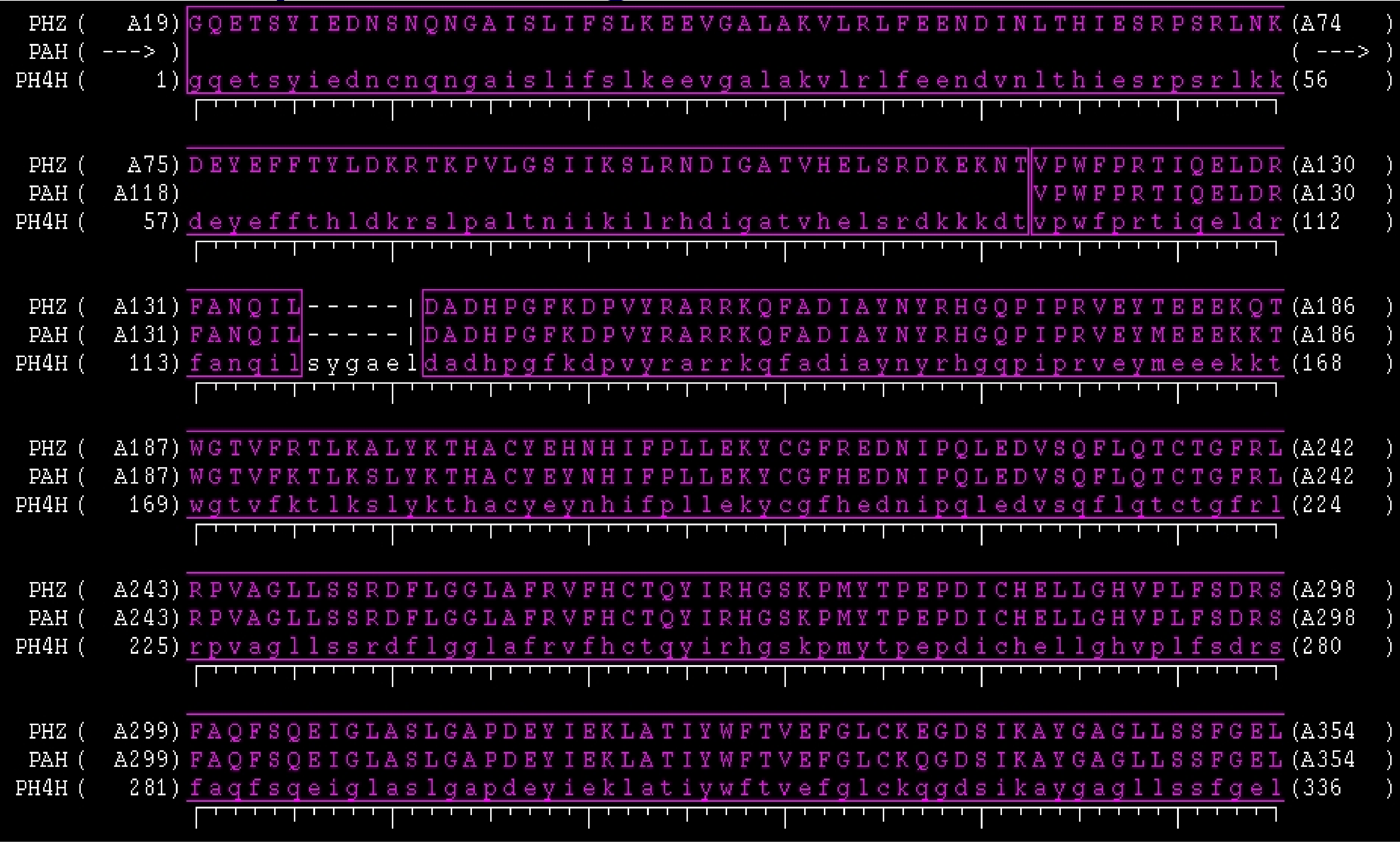
2PAH & 1PHZ



Superimpose the overlap region

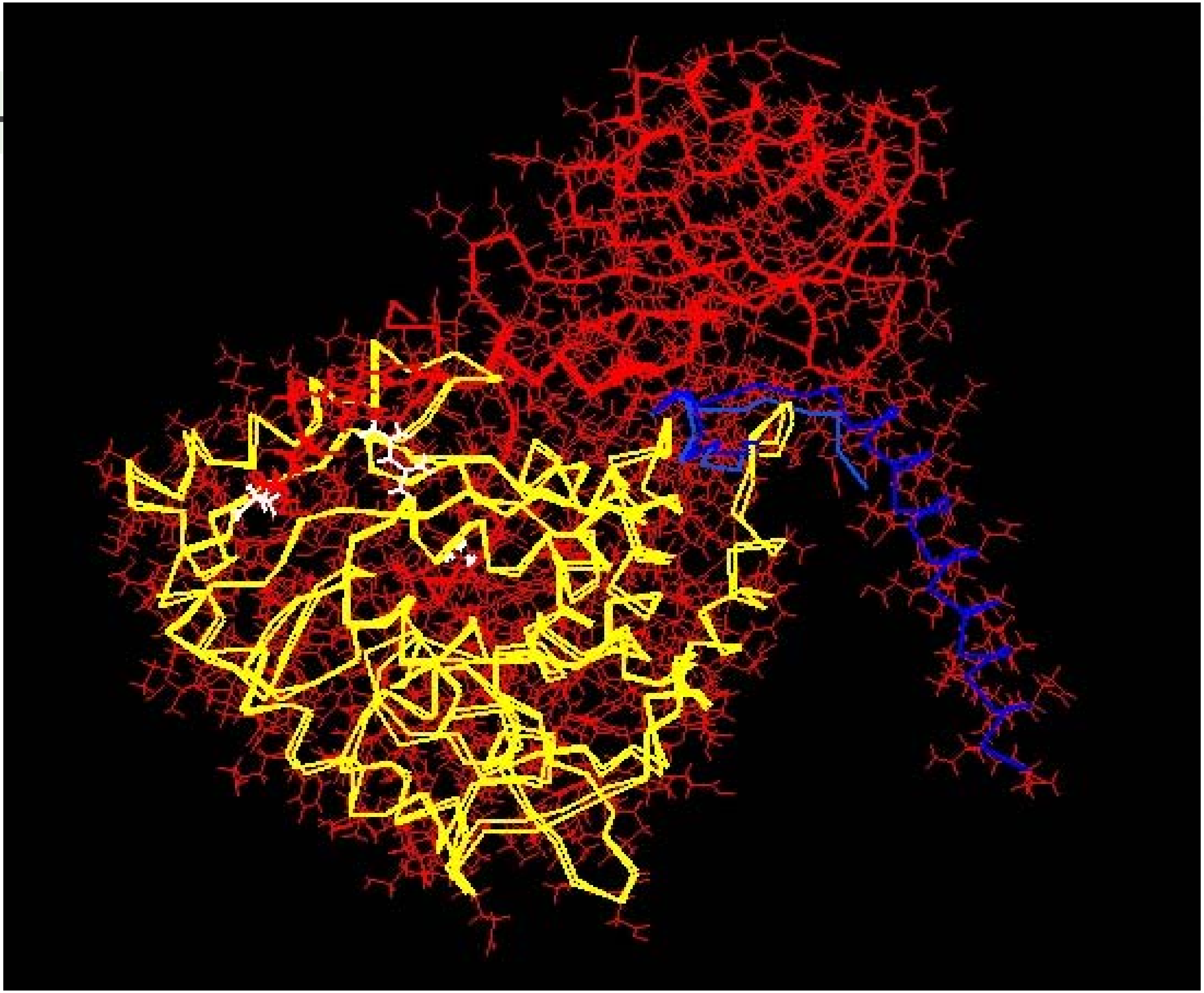


Sequence alignment



```
 PHZ (   A19) GQETSYIEDNSNQNGAISLIFSLKEEVGALAKVLRLFEENDINLTHIESRPSRLNK (A74   )
 PAH ( ---> )                                                          ( ---> )
PH4H (     1) gqetsyiedncnqngaislifslkeevgalakvlrlfeendvnlthiesrpsrlkk (56    )

 PHZ (   A75) DEYEFFTYLDKRTKPVLGSIIKSLRNDIGATVHELSRDKEKNTVPWFPRTIQELDR (A130  )
 PAH (  A118)                                            VPWFPRTIQELDR (A130  )
PH4H (    57) deyeffthldkrslpaltniikilrhdigatvhelsrdkkkdtvpwfprtiqeldr (112   )

 PHZ (  A131) FANQIL-----|DADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRVEYTEEEKQT (A186  )
 PAH (  A131) FANQIL-----|DADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRVEYMEEEKKT (A186  )
PH4H (   113) fanqilsygaeldadhpgfkdpvyrarrkqfadiaynyrhgqpiprveymeeekkt (168   )

 PHZ (  A187) WGTVFRTLKALYKTHACYEHNHIFPLLEKYCGFREDNIPQLEDVSQFLQTCTGFRL (A242  )
 PAH (  A187) WGTVFKTLKSLYKTHACYEYNHIFPLLEKYCGFHEDNIPQLEDVSQFLQTCTGFRL (A242  )
PH4H (   169) wgtvfktlkslykthacyeynhifpllekycgfhednipqledvsqflqtctgfrl (224   )

 PHZ (  A243) RPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDRS (A298  )
 PAH (  A243) RPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDRS (A298  )
PH4H (   225) rpvagllssrdflgglafrvfhctqyirhgskpmytpepdichellghvplfsdrs (280   )

 PHZ (  A299) FAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKEGDSIKAYGAGLLSSFGEL (A354  )
 PAH (  A299) FAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKQGDSIKAYGAGLLSSFGEL (A354  )
PH4H (   281) faqfsqeiglaslgapdeyieklatiywftvefglckqgdsikaygagllssfgel (336   )
```

In the process of model generating



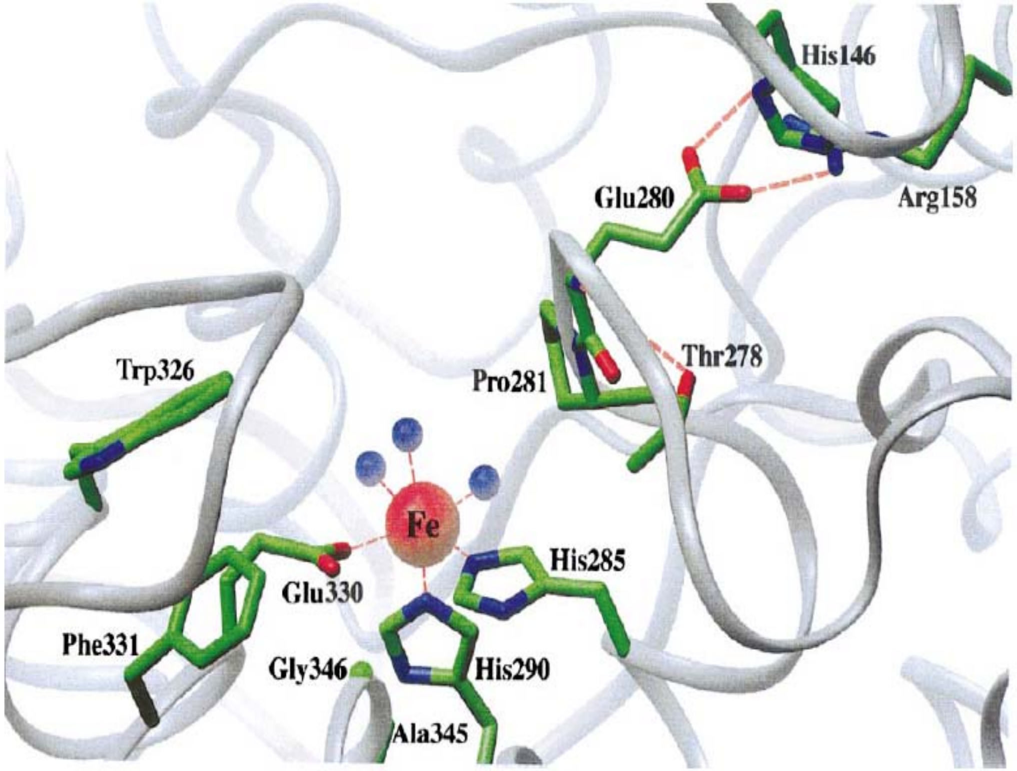
RMSD to its templates:

- To 1PHZ:
  - C-alpha RMSD: 0.67
- To 2PAH:
  - C-alpha RMSD: 0.89

- Note: these RMSD are calculated by the correspondingly aligned residues.

An example of the iron ion site of PAH analysis

- Actually this analysis doesn’t require the tetramer model; only those analysis in the interfaces between monomers require the tetramers.



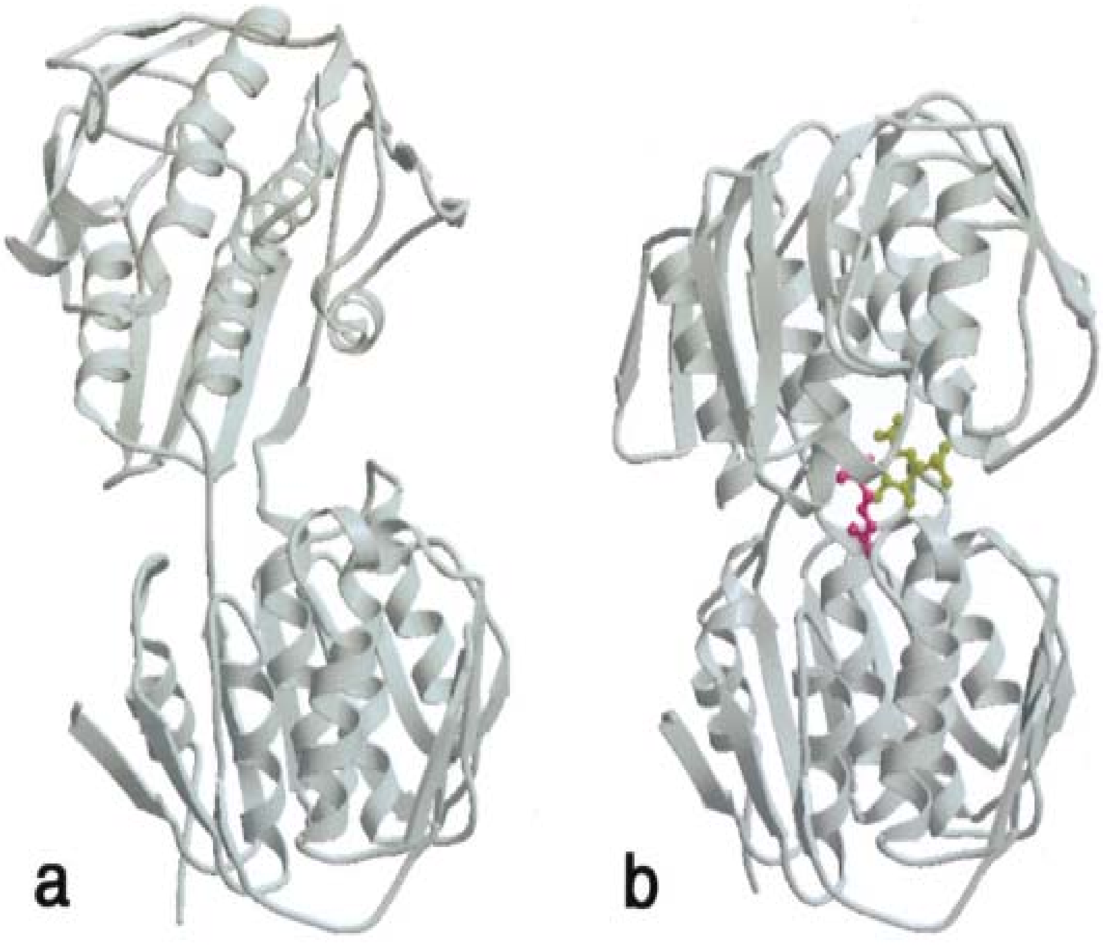
Modeling of rice EPSP synthase

- Background of EPSPS
- Search for the template of rice EPSPS
- Align the target and the template
- Build model
- Optimization
- Model check
- Analysis

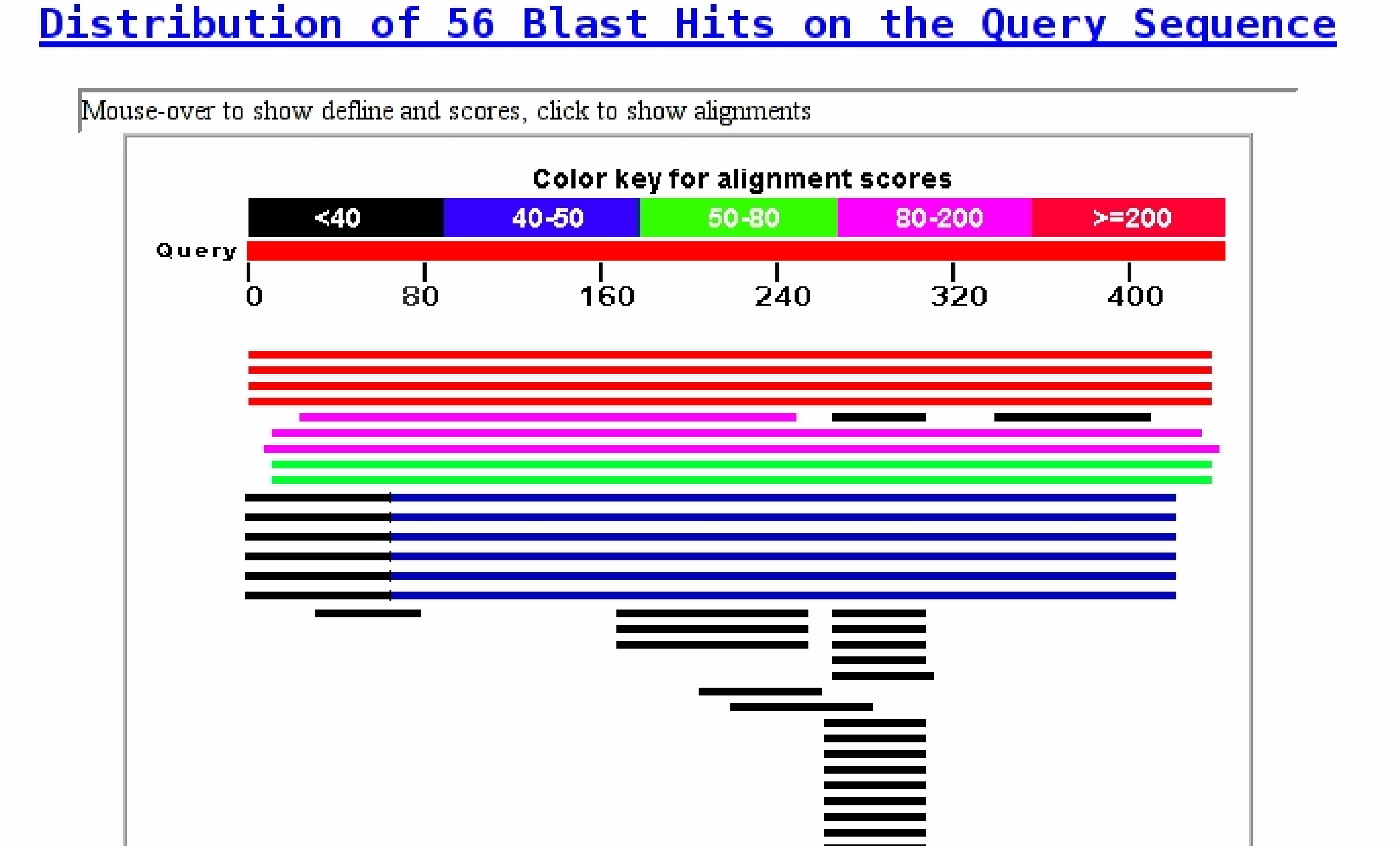
Background of EPSPS

- A key enzyme in the shikamate pathway, which is essential for plants and micro organisms, but is lack in the animals
- Catalysis: PEP + S3P → EPSP
- Glyphosate is an inhibitor, whose structure is similar to PEP.
- Thus, Glyphosate is used as herbicide. (How?)

A two-domain architecture

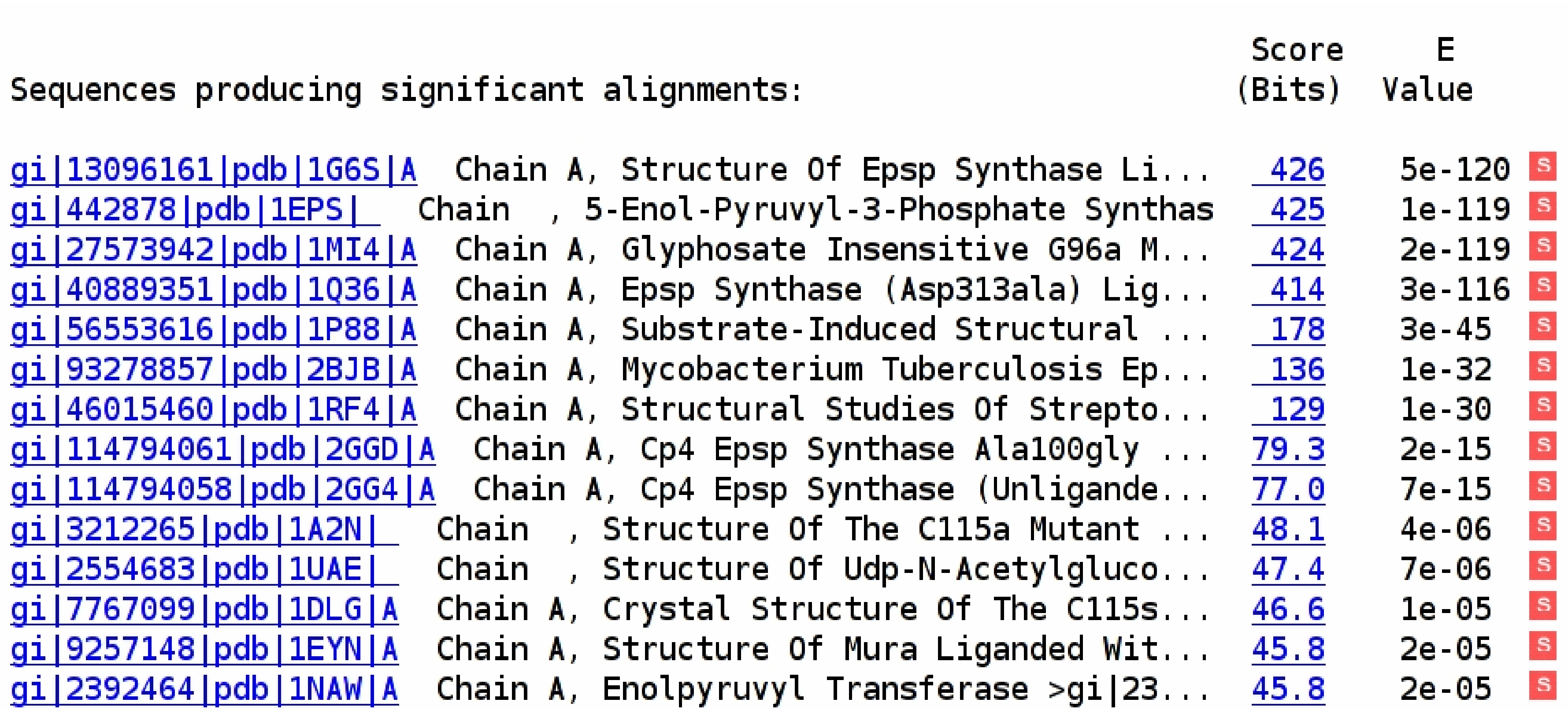


Search for the template of rice EPSPS

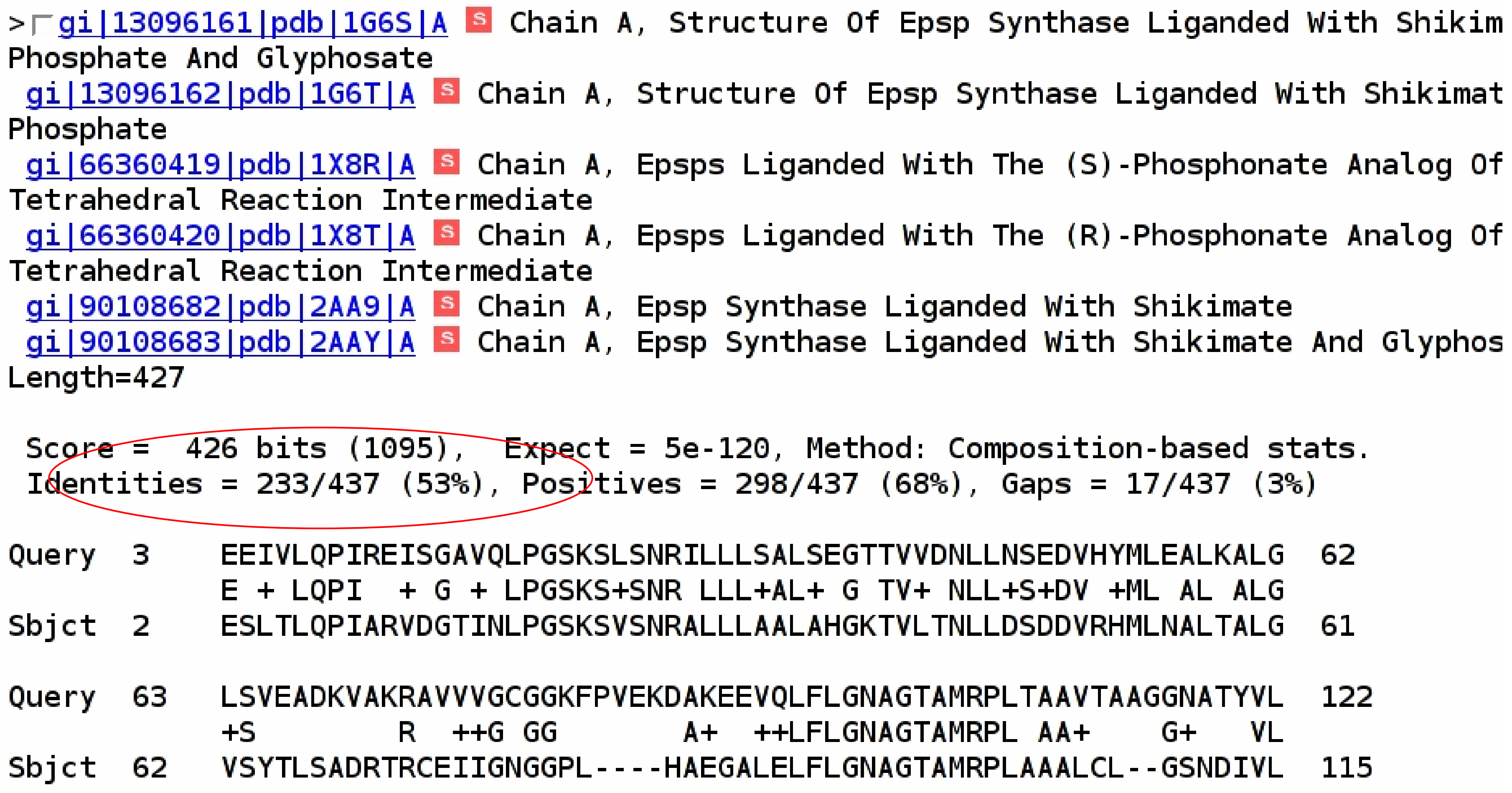


Distribution of 56 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

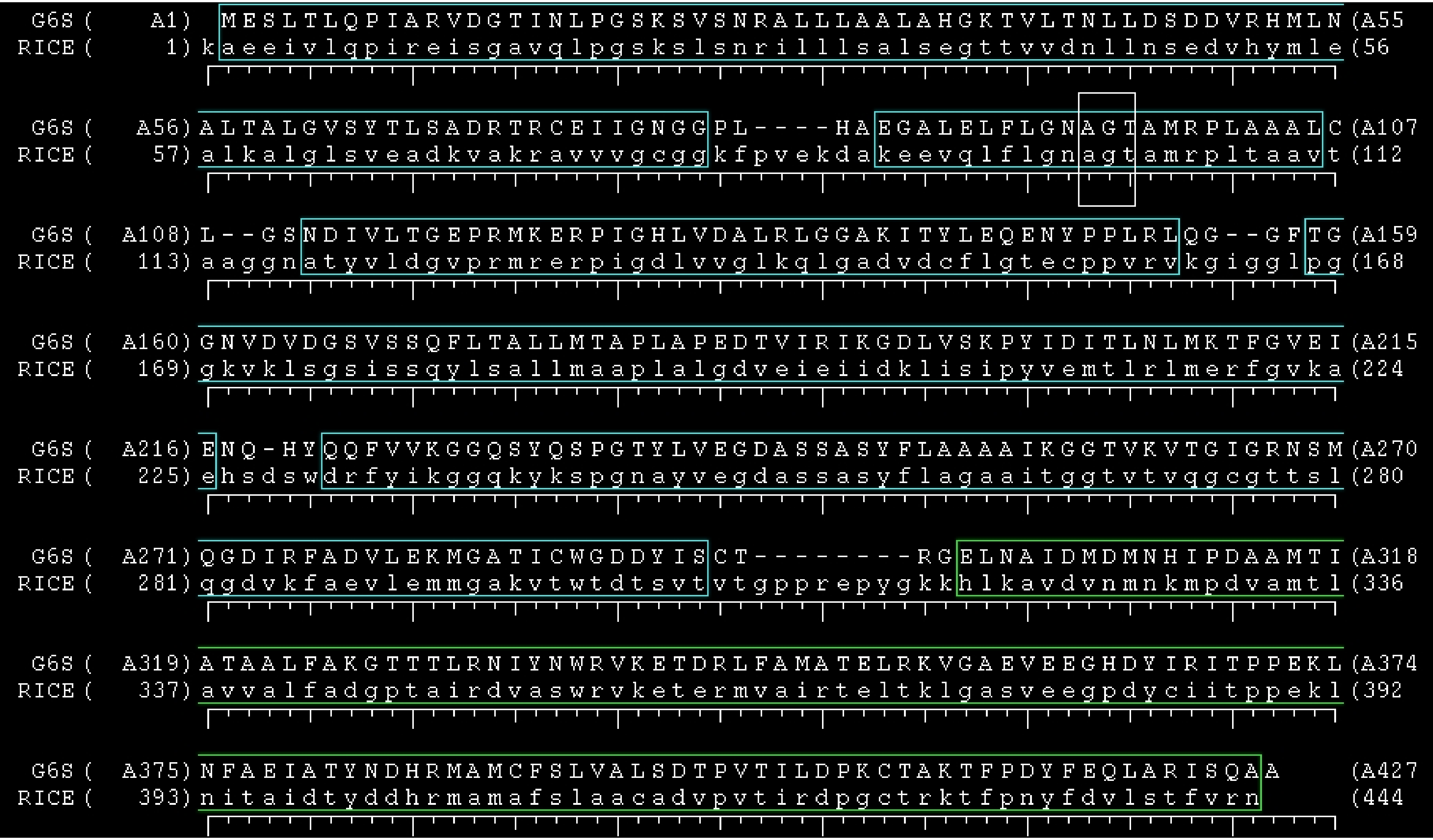


```
                                                                 Score    E
Sequences producing significant alignments:                      (Bits)  Value

gi|13096161|pdb|1G6S|A  Chain A, Structure Of Epsp Synthase Li...  426    5e-120 S
gi|442878|pdb|1EPS|   Chain  , 5-Enol-Pyruvyl-3-Phosphate Synthas  425    1e-119 S
gi|27573942|pdb|1MI4|A  Chain A, Glyphosate Insensitive G96a M...  424    2e-119 S
gi|40889351|pdb|1Q36|A  Chain A, Epsp Synthase (Asp313ala) Lig...  414    3e-116 S
gi|56553616|pdb|1P88|A  Chain A, Substrate-Induced Structural ...  178    3e-45  S
gi|93278857|pdb|2BJB|A  Chain A, Mycobacterium Tuberculosis Ep...  136    1e-32  S
gi|46015460|pdb|1RF4|A  Chain A, Structural Studies Of Strepto...  129    1e-30  S
gi|114794061|pdb|2GGD|A  Chain A, Cp4 Epsp Synthase Ala100gly ...  79.3   2e-15  S
gi|114794058|pdb|2GG4|A  Chain A, Cp4 Epsp Synthase (Unligande...  77.0   7e-15  S
gi|3212265|pdb|1A2N|   Chain  , Structure Of The C115a Mutant ...  48.1   4e-06  S
gi|2554683|pdb|1UAE|   Chain  , Structure Of Udp-N-Acetylgluco...  47.4   7e-06  S
gi|7767099|pdb|1DLG|A  Chain A, Crystal Structure Of The C115s...  46.6   1e-05  S
gi|9257148|pdb|1EYN|A  Chain A, Structure Of Mura Liganded Wit...  45.8   2e-05  S
gi|2392464|pdb|1NAW|A  Chain A, Enolpyruvyl Transferase >gi|23...  45.8   2e-05  S
```



```
>gi|13096161|pdb|1G6S|A  Chain A, Structure Of Epsp Synthase Liganded With Shikim
Phosphate And Glyphosate
 gi|13096162|pdb|1G6T|A  Chain A, Structure Of Epsp Synthase Liganded With Shikimat
Phosphate
 gi|66360419|pdb|1X8R|A  Chain A, Epsps Liganded With The (S)-Phosphonate Analog Of
Tetrahedral Reaction Intermediate
 gi|66360420|pdb|1X8T|A  Chain A, Epsps Liganded With The (R)-Phosphonate Analog Of
Tetrahedral Reaction Intermediate
 gi|90108682|pdb|2AA9|A  Chain A, Epsp Synthase Liganded With Shikimate
 gi|90108683|pdb|2AAY|A  Chain A, Epsp Synthase Liganded With Shikimate And Glyphos
Length=427

 Score =  426 bits (1095),  Expect = 5e-120, Method: Composition-based stats.
 Identities = 233/437 (53%), Positives = 298/437 (68%), Gaps = 17/437 (3%)

Query  3    EEIVLQPIREISGAVQLPGSKSLSNRILLLSALSEGTTVVDNLLNSEDVHYMLEALKALG  62
            E + LQPI  + G + LPGSKS+SNR LLL+AL+ G TV+ NLL+S+DV +ML AL ALG
Sbjct  2    ESLTLQPIARVDGTINLPGSKSVSNRALLLAALAHGKTVLTNLLDSDDVRHMLNALTALG  61

Query  63   LSVEADKVAKRAVVVGCGGKFPVEKDAKEEVQLFLGNAGTAMRPLTAAVTAAGGNATYVL  122
            +S        R  ++G GG     A+  ++LFLGNAGTAMRPL AA+    G+   VL
Sbjct  62   VSYTLSADRTRCEIIGNGGPL----HAEGALELFLGNAGTAMRPLAAALCL--GSNDIVL  115
```

One possible alignment



```
 G6S (    A1)  MESLTLQPIARVDGTINLPGSKSVSNRALLLAALAHGKTVLTNLLDSDDVRHMLN (A55
RICE (     1) kaeeivlqpireisgavqlpgskslsnrillllsalsegttvvdnllnsedvhymle (56

 G6S (   A56) ALTALGVSYTLSADRTRCEIIGNGGPL----HAEGALELFLGNAGTAMRPLAAALC (A107
RICE (    57) alkalglsveadkvakravvvgcggkfpvekdakeevqlflgnagtamrpltaavt (112

 G6S (  A108) L--GSNDIVLTGEPRMKERPIGHLVDALRLGGAKITYLEQENYPPLRLQG--GFTG (A159
RICE (   113) aaggnatyvldgvprmrerpigdlvvglkqlgadvdcflgtecppvrvkgigglpg (168

 G6S (  A160) GNVDVDGSVSSQFLTALLMTAPLAPEDTVIRIKGDLVSKPYIDITLNLMKTFGVEI (A215
RICE (   169) gkvklsgsissqylsallmaaplalgdveieiidklisipyvemtlrlmerfgvka (224

 G6S (  A216) ENQ-HYQQFVVKGGQSYQSPGTYLVEGDASSASYFLAAAAIKGGTVKVTGIGRNSM (A270
RICE (   225) ehsdswdrfyikggqkykspgnayvegdassasyflagaaitggtvtvqgcgttsl (280

 G6S (  A271) QGDIRFADVLEKMGATICWGDDYISCT--------RGELNAIDMDMNHIPDAAMTI (A318
RICE (   281) qgdvkfaevlemmgakvtwtdtsvtvtgpprepygkkhlkavdvnmnkmpdvamtl (336

 G6S (  A319) ATAALFAKGTTTLRNIYNWRVKETDRLFAMATELRKVGAEVEEGHDYIRITPPEKL (A374
RICE (   337) avvalfadgptairdvaswrvketermvairteltklgasveegpdyciitppekl (392

 G6S (  A375) NFAEIATYNDHRMAMCFSLVALSDTPVTILDPKCTAKTFPDYFEQLARISQAA    (A427
RICE (   393) nitaidtyddhrmamafslaacadvpvtirdpgctrktfpnyfdvlstfvrn     (444
```

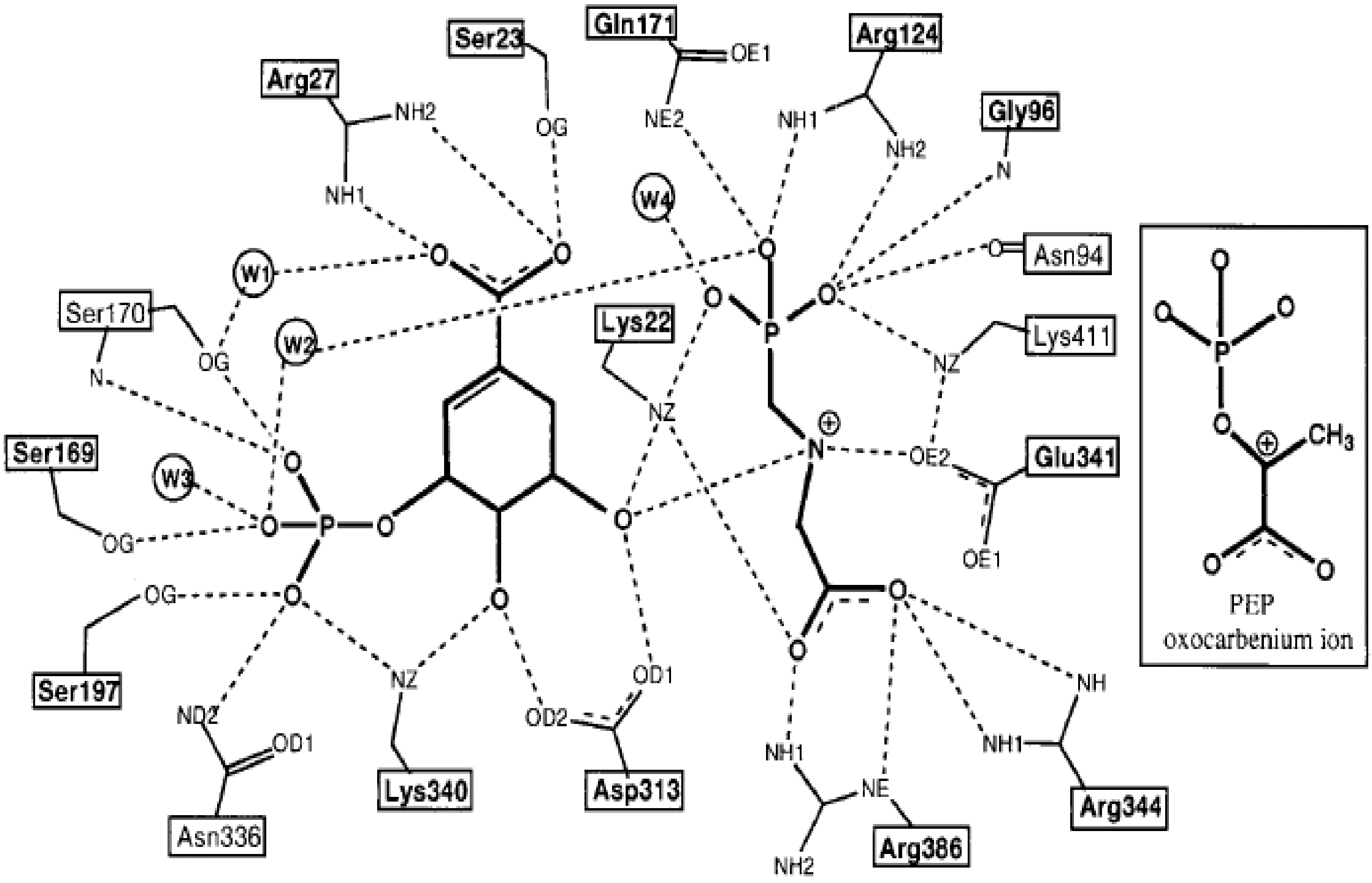
RMSD compared to the template

- After generating the model and the optimization, we can compare the C-alpha RMSD between the aligned residue pairs:
- RMSD = 0.149

Model check

- Procheck:
  - http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

The active site of *E.coli* EPSPS



The most important is the analysis

- Structural model is just a model
- Combining your experimental data to generate rational mechanism explanation is the most important

Additional slides

- Structural genomics
- CASP
- CAPRI

Thanks!