Supplementary data for article:

Petrović, D.; Zlatović, M. Modeling Human Serum Albumin Tertiary Structure to Teach Upper-Division Chemistry Students Bioinformatics and Homology Modeling Basics. *Journal of Chemical Education* **2015**, *92* (7), 1233–1237. https://doi.org/10.1021/ed500358f

Modeling Human Serum Albumin Tertiary Structure To Teach Upper-Division Chemistry Students Bioinformatics and Homology Modeling Basics (Step-By-Step Lab Manual)

Dušan Petrović and Mario Zlatović

Faculty of Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia

CONTENTS:

INTRODUCTION

In the present laboratory experiment homology modeling will be performed. This method is being used when the crystal structure of the protein of interest is not known, but it is necessary for further modeling. Homology models can be used to study dynamics of protein or to design a ligand or matrix for affinity chromatography protein purification.

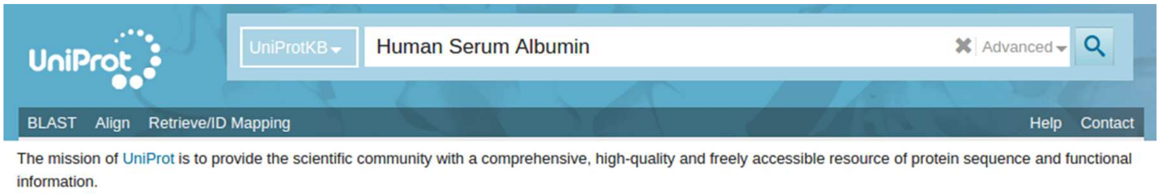
Crystal structure of the human serum albumin (HSA) was determined experimentally. However, in the present experiment a homology model of HSA will be prepared based on a serum albumin from another animal. The whole procedure will be done as if the HSA structure had not been known.

At the beginning, search for the HSA amino acid sequence and for the sequences of homologous proteins will be made. Once they are found, sequence alignment has to be performed in order to determine which protein with known crystal structure is the most suitable template. After making the choice, a template will be prepared for homology modeling, and homology model will be built. The prepared HSA homology model will be analyzed with several on-line tools.

At the end, quality of the prepared homology model will be benchmarked against PDB deposited crystal structures of HSA itself.

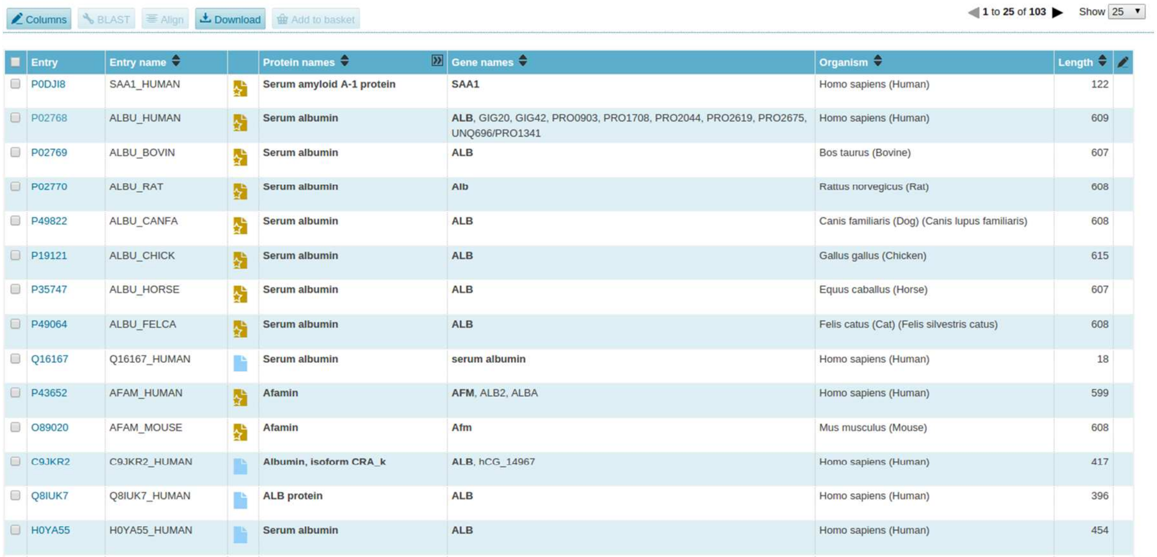
PART I: HSA SEQUENCE DATA MINING

In order to build a homology model of a protein of interest (POI), one needs to know POI's amino acid sequence. UniProt (http://www.uniprot.org/) is one of the top databases for protein sequence and functional information. To find the sequence of our POI, query for the "Human Serum Albumin" at the UniProt website and hit the magnifier button to search.



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

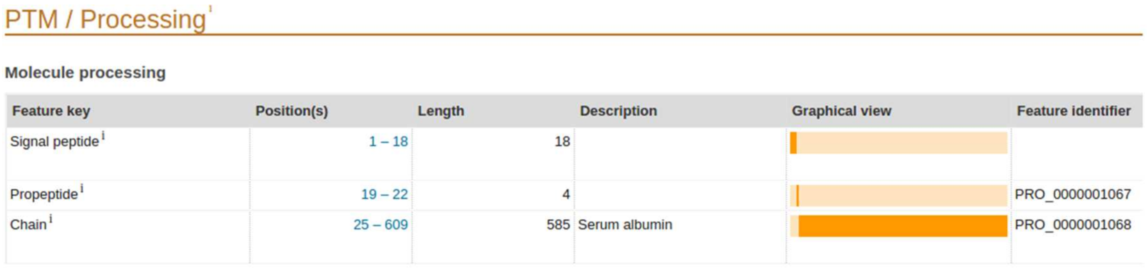
There are 103 listed results for this query. Take care that not all proteins from the list are actually serum albumins. Also, take care of the organism, since beside *Homo sapiens* there are other organisms as well. Finally, sequences with the blue paper sign are not reviewed, so whenever entry represented by the gold paper with a star is available it is recommended to use this, reviewed sequence.



☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	P0DJI8	SAA1_HUMAN	★	**Serum amyloid A-1 protein**	**SAA1**	Homo sapiens (Human)	122
☐	P02768	ALBU_HUMAN	★	**Serum albumin**	**ALB**, GIG20, GIG42, PRO0903, PRO1708, PRO2044, PRO2619, PRO2675, UNQ696/PRO1341	Homo sapiens (Human)	609
☐	P02769	ALBU_BOVIN	★	**Serum albumin**	**ALB**	Bos taurus (Bovine)	607
☐	P02770	ALBU_RAT	★	**Serum albumin**	**Alb**	Rattus norvegicus (Rat)	608
☐	P49822	ALBU_CANFA	★	**Serum albumin**	**ALB**	Canis familiaris (Dog) (Canis lupus familiaris)	608
☐	P19121	ALBU_CHICK	★	**Serum albumin**	**ALB**	Gallus gallus (Chicken)	615
☐	P35747	ALBU_HORSE	★	**Serum albumin**	**ALB**	Equus caballus (Horse)	607
☐	P49064	ALBU_FELCA	★	**Serum albumin**	**ALB**	Felis catus (Cat) (Felis silvestris catus)	608
☐	Q16167	Q16167_HUMAN	📄	**Serum albumin**	**serum albumin**	Homo sapiens (Human)	18
☐	P43652	AFAM_HUMAN	★	**Afamin**	**AFM**, ALB2, ALBA	Homo sapiens (Human)	599
☐	O89020	AFAM_MOUSE	★	**Afamin**	**Afm**	Mus musculus (Mouse)	608
☐	C9JKR2	C9JKR2_HUMAN	📄	**Albumin, isoform CRA_k**	**ALB**, hCG_14967	Homo sapiens (Human)	417
☐	Q8IUK7	Q8IUK7_HUMAN	📄	**ALB protein**	**ALB**	Homo sapiens (Human)	396
☐	H0YA55	H0YA55_HUMAN	📄	**Serum albumin**	**ALB**	Homo sapiens (Human)	454

Based on the Results page, P02768 entry should be chosen: it is a serum albumin from *Homo sapiens* and it is a reviewed sequence. Open the sequence by clicking on the entry ID. Numerous information about HAS is given, but the "PTM / Processing" section is the most important for this experiment.

The complete protein sequence is 609 amino acids long. The signal peptide is from amino acid 1 to 18 (18 residues), while propeptide is from amino acid 19 to 22 (4 residues). Both these peptides have to be cleaved to produce mature protein. In the future modeling only the mature HSA, from amino acid 25 to 609 (585 residues) will be used. To obtain its sequence, in the "Molecule processing" subsection click at the orange bar (corresponding to the "Chain") in the graphical view column (feature identifier: PRO_0000001068).



PTM / Processing

Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Signal peptide	1 – 18	18			
Propeptide	19 – 22	4			PRO_0000001067
Chain	25 – 609	585	Serum albumin		PRO_0000001068

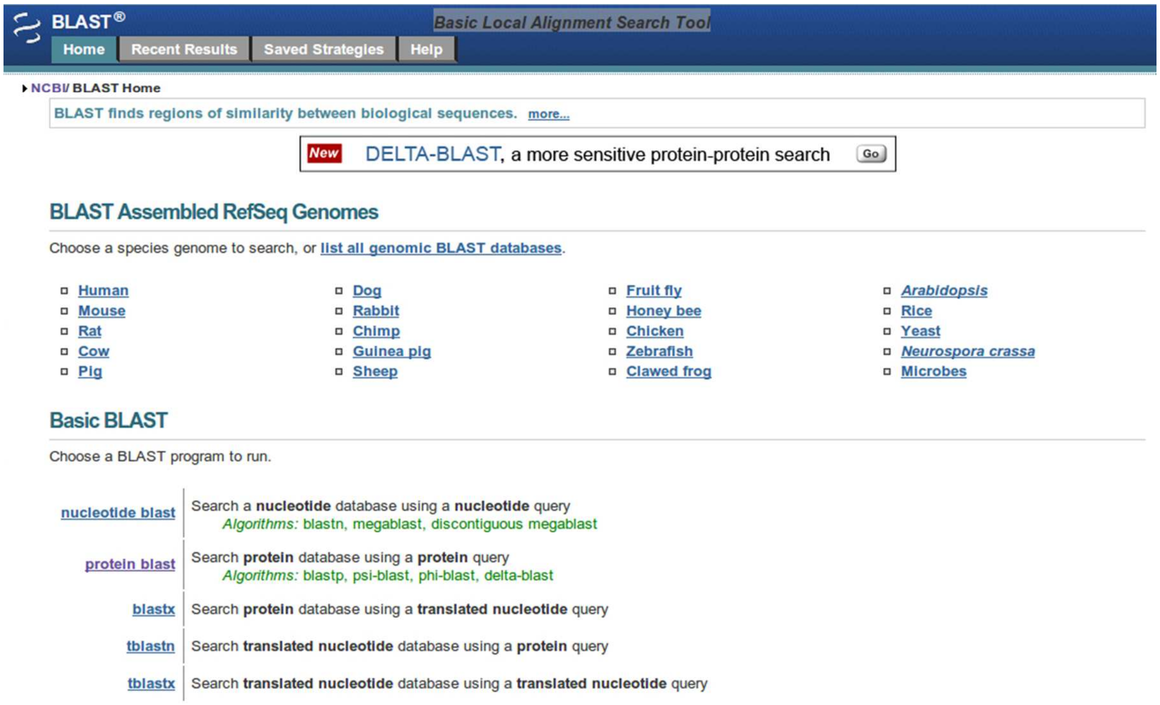
The following sequence in FASTA format appears:

```
>HSA
DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE
NCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEV
DVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLP
KLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK
VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPA
DLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKC
CAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST
PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTES
LVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKAT
KEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

and exactly this sequence will be used for the following simulations.

PART II: FINDING PROTEINS HOMOLOGOUS TO HSA

After the sequence of POI is obtained, search for its homologous proteins should be made. One of the most frequently used tools for protein comparison is Basic Local Alignment Search Tool (BLAST), available at the http://blast.ncbi.nlm.nih.gov/ web page.



NCBI/ BLAST Home

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

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled RefSeq Genomes

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

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Sheep
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- *Neurospora crassa*
- Microbes

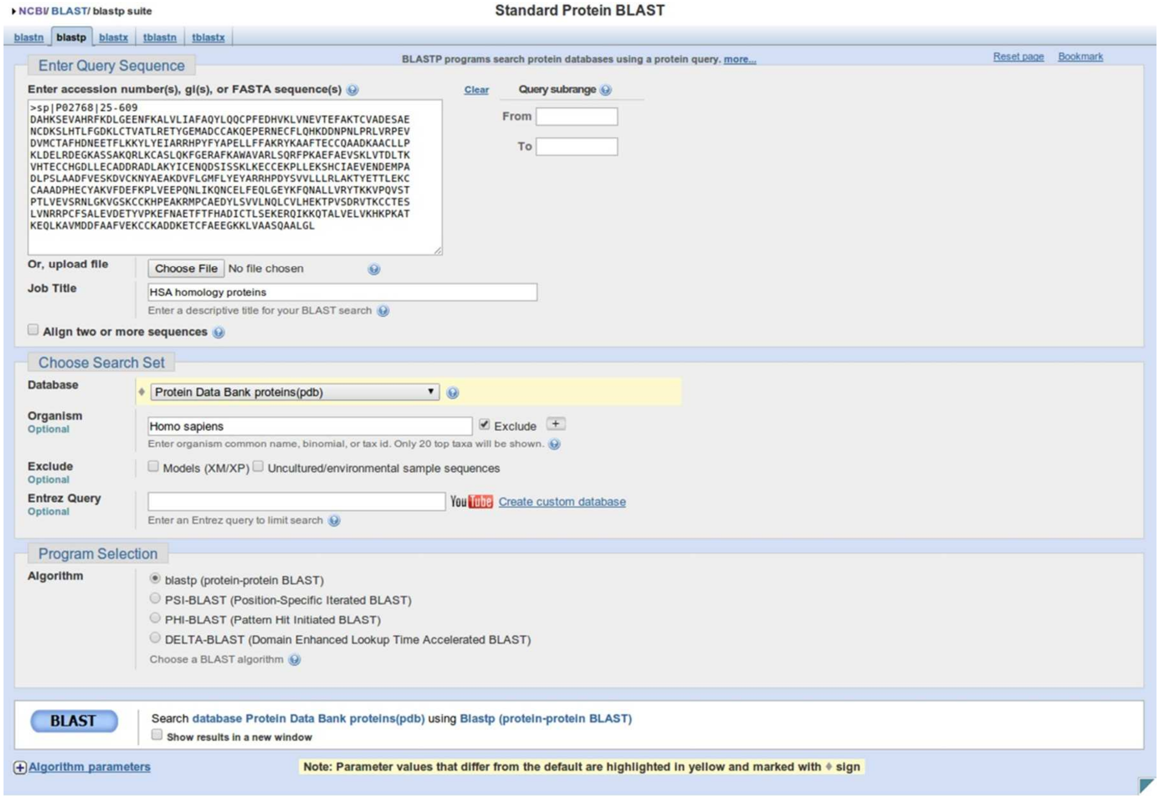
Basic BLAST

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, delta-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

In order to compare proteins, from the “Basic BLAST” section choose “protein blast” command.

In the “Enter Query Sequence” section paste the FASTA formatted sequence of the mature HSA, and enter descriptive “Job Title”. From the “Choose Search Set” section choose “Protein Data Bank proteins(pdb)” database. With this option, searching for only those proteins whose tertiary structures are known will be made. In the same section exclude “Homo sapiens” organism; otherwise majority of the found sequences would be the same HSA. In the “Program Selection” section choose “blastp (protein-protein BLAST)” algorithm, and finally click the *BLAST* button.



▸ NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P02768|25-609
DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE
NCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEV
DVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLP
KLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK
VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPA
DLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKC
CAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST
PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTES
LVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKAT
KEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

Clear

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: HSA homology proteins

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Protein Data Bank proteins(pdb)

Organism (Optional): Homo sapiens ☑ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query (Optional):

Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

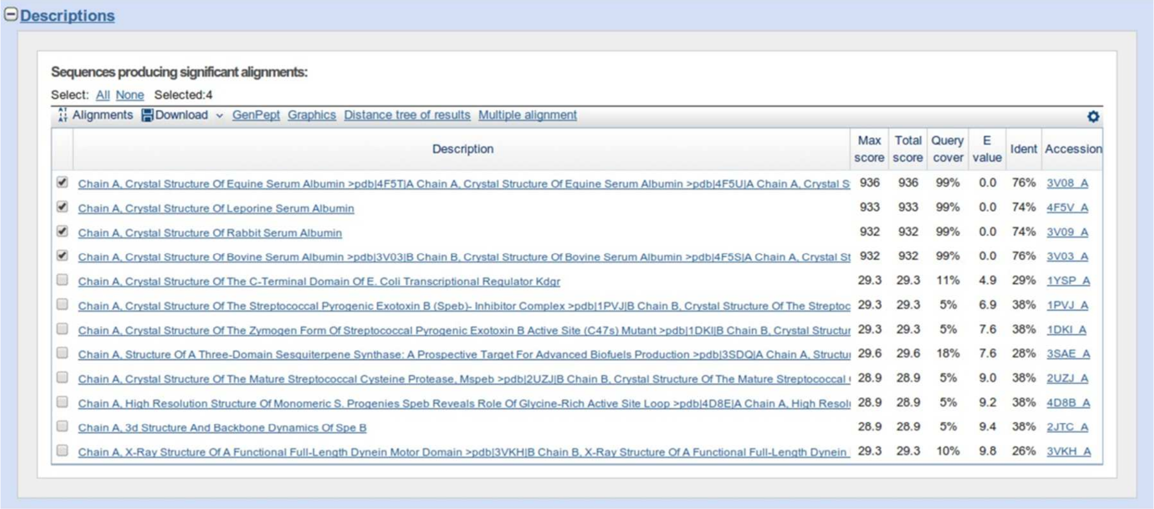
Search database Protein Data Bank proteins(pdb) using Blastp (protein-protein BLAST)

☐ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

BLAST search gives 12 possible results, but only first 4 (checked) have identity with HSA over 50%, and their E-value is zero. The expectation (E) value represents the number of different alignments equivalent to or better than the alignment that is expected to occur in a database search by chance. Practically, the lower the E-value, the better the alignment. As the rule of thumb, sequences with E-value > 1 are not suitable for homology modeling. The last 8 hits will not be further considered due to the very small homology percentage and high E-values.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:4

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☑	Chain A, Crystal Structure Of Equine Serum Albumin >pdb|4F5T|A Chain A, Crystal Structure Of Equine Serum Albumin >pdb|4F5U|A Chain A, Crystal S	936	936	99%	0.0	76%	3V08_A
☑	Chain A, Crystal Structure Of Leporine Serum Albumin	933	933	99%	0.0	74%	4F5V_A
☑	Chain A, Crystal Structure Of Rabbit Serum Albumin	932	932	99%	0.0	74%	3V09_A
☑	Chain A, Crystal Structure Of Bovine Serum Albumin >pdb|3V03|B Chain B, Crystal Structure Of Bovine Serum Albumin >pdb|4F5S|A Chain A, Crystal St	932	932	99%	0.0	76%	3V03_A
☐	Chain A, Crystal Structure Of The C-Terminal Domain Of E. Coli Transcriptional Regulator Kdgr	29.3	29.3	11%	4.9	29%	1YSP_A
☐	Chain A, Crystal Structure Of The Streptococcal Pyrogenic Exotoxin B (Speb)- Inhibitor Complex >pdb|1PVJ|B Chain B, Crystal Structure Of The Streptoc	29.3	29.3	5%	6.9	38%	1PVJ_A
☐	Chain A, Crystal Structure Of The Zymogen Form Of Streptococcal Pyrogenic Exotoxin B Active Site (C47s) Mutant >pdb|1DKI|B Chain B, Crystal Structur	29.3	29.3	5%	7.6	38%	1DKI_A
☐	Chain A, Structure Of A Three-Domain Sesquiterpene Synthase: A Prospective Target For Advanced Biofuels Production >pdb|3SDQ|A Chain A, Structui	29.6	29.6	18%	7.6	28%	3SAE_A
☐	Chain A, Crystal Structure Of The Mature Streptococcal Cysteine Protease, Mspeb >pdb|2UZJ|B Chain B, Crystal Structure Of The Mature Streptococcal (	28.9	28.9	5%	9.0	38%	2UZJ_A
☐	Chain A, High Resolution Structure Of Monomeric S. Progenies Speb Reveals Role Of Glycine-Rich Active Site Loop >pdb|4D8E|A Chain A, High Resoli	28.9	28.9	5%	9.2	38%	4D8B_A
☐	Chain A, 3d Structure And Backbone Dynamics Of Spe B	28.9	28.9	5%	9.4	38%	2JTC_A
☐	Chain A, X-Ray Structure Of A Functional Full-Length Dynein Motor Domain >pdb|3VKH|B Chain B, X-Ray Structure Of A Functional Full-Length Dynein	29.3	29.3	10%	9.8	26%	3VKH_A

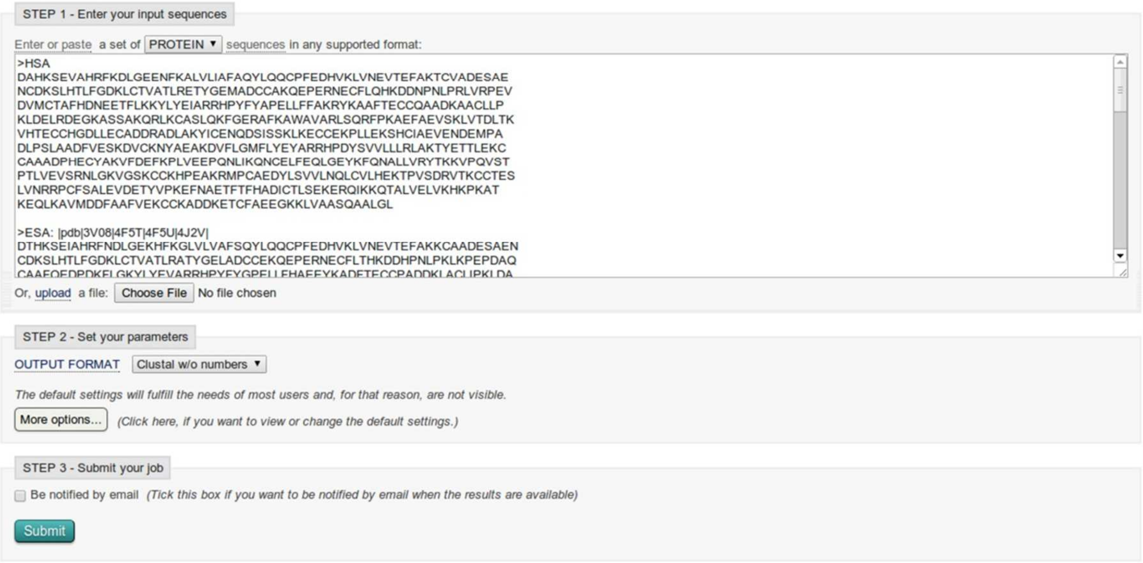
To download sequences of the four checked structures, click on the *Download* button and check “FASTA (complete sequence)”. Obtained sequences are the following:

```
>ESA: |pdb|3V08|4F5T|4F5U|4J2V|
DTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPFEDHVKLVNEVTEFAKKCAADESAEN
CDKSLHTLFGDKLCTVATLRATYGELADCCEKQEPERNECFLTHKDDHPNLPKLKPEPDAQ
CAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFHAEEYKADFTECCPADDKLACLIPKLDA
LKERILLSSAKERLKCSSFQNFGERAVKAWSVARLSQKFPKADFAEVSKIVTDLTKVHKEC
CHGDLLECADDRADLAKYICEHQDSISGKLKACCDKPLLQKSHCIAEVKEDDLPSDLPALA
ADFAEDKEICKHYKDAKDVFLGTFLYEYSRRHPDYSVSLLLRIAKTYEATLEKCCAEADPP
ACYRTVFDQFTPLVEEPKSLVKKNCDLFEEVGEYDFQNALIVRYTKKAPQVSTPTLVEIGR
TLGKVGSRCCKLPESERLPCSENHLALALNRLCVLHEKTPVSEKITKCCTDSLAERRPCFS
ALELDEGYVPKEFKAETFTFHADICTLPEDEKQIKKQSALAELVKHKPKATKEQLKTVLGN
FSAFVAKCCGR EDKEACFAEEGPKLVASSQLALA

>LSA: |pdb|4F5V|
EAHKSEIAHRFNDVGEEHFIGLVLITFSQYLQKCPYEEHAKLVKEVTDLAKACVADESAAN
CDKSLHDIFGDKICALPSLRDTYGDVADCCEKKEPERNECFLHHKDDKPDLPPFARPEADV
LCKAFHDDEKAFFGHYLYEVARRHPYFYAPELLYYAQKYKAILTECCEAADKGACLTPKLD
ALKEKALISAAQERLRCASIQKFGDRAYKAWALVRLSQRFPKADFTDISKIVTDLTKVHKE
CCHGDLLECADDRADLAKYMCEHQETISSHLKECCDKPILEKAHCIYGLHNDETPAGLPAV
AEEFVEDKDVCKNYEEAKDLFLGKFLYEYSRRHPDYSVVLLLRLGKAYEATLKKCCATDDP
HACYAKVLDEFQPLVDEPKNLVKQNCELYEQLGDYNFQNALLVRYTKKVPQVSTPTLVEIS
RSLGKVGSKCCKHPEAERLPCVEDYLSVVLNRLCVLHEKTPVSEKVTKCCSESLVDRRPCF
SALGPDETYVPKEFNAETFTFHADICTLPETERKIKKQTALVELVKHKPHATNDQLKTVVG
EFTALLDKCCS AEDKEACFAVEGPKLVESSKATLG

>RSA: |pdb|3V09|
EAHKSEIAHRFNDVGEEHFIGLVLITFSQYLQKCPYEEHAKLVKEVTDLAKACVADESAAN
CDKSLHDIFGDKICALPSLRDTYGDVADCCEKKEPERNECFLHHKDDKPDLPPFARPEADV
LCKAFHDDEKAFFGHYLYEVARRHPYFYAPELLYYAQKYKAILTECCEAADKGACLTPKLD
ALEGKSLISAAQERLRCASIQKFGDRAYKAWALVRLSQRFPKADFTDISKIVTDLTKVHKE
CCHGDLLECADDRADLAKYMCEHQETISSHLKECCDKPILEKAHCIYGLHNDETPAGLPAV
AEEFVEDKDVCKNYEEAKDLFLGKFLYEYSRRHPDYSVVLLLRLGKAYEATLKKCCATDDP
HACYAKVLDEFQPLVDEPKNLVKQNCELYEQLGDYNFQNALLVRYTKKVPQVSTPTLVEIS
RSLGKVGSKCCKHPEAERLPCVEDYLSVVLNRLCVLHEKTPVSEKVTKCCSESLVDRRPCF
SALGPDETYVPKEFNAETFTFHADICTLPETERKIKKQTALVELVKHKPHATNDQLKTVVG
EFTALLDKCCS AEDKEACFAVEGPKLVESSKATLG

>BSA: |pdb|3V03|4F5S|4JK4|
DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAG
CEKSLHTLFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPKLKPDPNTL
CDEFKADEKKFWGKYLYEIARRHPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIET
MREKVLTSSARQRLRCASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKEC
CHGDLLECADDRADLAKYICDNQDTISSKLKECCDKPLLEKSHCIAEVEKDAIPENLPPLT
ADFAEDKDVCKNYQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAKDDPH
ACYSTVFDKLKHLVDEPQNLIKQNCDQFEKLGEYGFQNALIVRYTRKVPQVSTPTLVEVSR
SLGKVGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTESLVNRRPCFS
ALTPDETYVPKAFDEKLFTFHADICTLPDTEKQIKKQTALVELLKHKPKATEEQLKTVMEN
FVAFVDKCCAA DDKEACFAVEGPKLVVSTQTALA
```

where **ESA** stands for Equine Serum Albumin, **LSA** stands for Leporine Serum Albumin, **RSA** stand for Rabbit Serum Albumin and **BSA** stands for Bovine Serum Albumin.

PART III: HSA, ESA, LSA, RSA & BSA SEQUENCE ALIGNMENT

Now the sequences of HSA homologous proteins should be aligned to inspect their similarity and differences. For sequence alignment Clustal Omega server will be used (https://www.ebi.ac.uk/Tools/msa/clustalo/). In the "STEP 1 - Enter your input sequences" section sequence of the HSA should be pasted, as well as the sequences of four homologous proteins (ESA, LSA, RSA and BSA), and then *Submit* button should be pressed.



STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN ▾ sequences in any supported format:

```
>HSA
DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE
NCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEV
DVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLP
KLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK
VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPA
DLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKC
CAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST
PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTES
LVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKAT
KEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL

>ESA: |pdb|3V08|4F5T|4F5U|4J2V|
DTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPFEDHVKLVNEVTEFAKKCAADESAEN
CDKSLHTLFGDKLCTVATLRATYGELADCCEKQEPERNECFLTHKDDHPNLPKLKPEPDAQ
CAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFHAEEYKADFTECCPADDKLACLIPKLDA
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your parameters

OUTPUT FORMAT Clustal w/o numbers ▾

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

The following sequence alignment appears:

```
HSA      DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE
ESA      DTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPFEDHVKLVNEVTEFAKKCAADESAE
LSA      EAHKSEIAHRFNDVGEEHFIGLVLITFSQYLQKCPYEEHAKLVKEVTDLAKACVADESAA
RSA      EAHKSEIAHRFNDVGEEHFIGLVLITFSQYLQKCPYEEHAKLVKEVTDLAKACVADESAA
BSA      DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHA
         ::****:****:*:**:.* .***::*:****:**:::*.***:*:*::** *.****

HSA      NCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEV
ESA      NCDKSLHTLFGDKLCTVATLRATYGELADCCEKQEPERNECFLTHKDDHPNLPKL-KPEP
LSA      NCDKSLHDIFGDKICALPSLRDTYGDVADCCEKKEPERNECFLHHKDDKPDLPPFARPEA
RSA      NCDKSLHDIFGDKICALPSLRDTYGDVADCCEKKEPERNECFLHHKDDKPDLPPFARPEA
BSA      GCEKSLHTLFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPKL-KPDP
          *:**** :***::* : :** ***::**** *:********* **** *:** : :*:

HSA      DVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLP
ESA      DAQCAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFHAEEYKADFTECCPADDKLACLIP
LSA      DVLCKAFHDDEKAFFGHYLYEVARRHPYFYAPELLYYAQKYKAILTECCEAADKGACLTP
RSA      DVLCKAFHDDEKAFFGHYLYEVARRHPYFYAPELLYYAQKYKAILTECCEAADKGACLTP
BSA      NTLCDEFKADEKKFWGKYLYEIARRHPYFYAPELLYYANKYNGVFQECCQAEDKGACLLP
         :. *  *: : . *  :****:********.****:.*:.*:. : *** * ** *** *
HSA      KLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK
ESA      KLDALKERILLSSAKERLKCSSFQNFGERAVKAWSVARLSQKFPKADFAEVSKIVTDLTK
```

```
LSA        KLDALKEKALISAAQERLRCASIQKFGDRAYKAWALVRLSQRFPKADFTDISKIVTDLTK
RSA        KLDALEGKSLISAAQERLRCASIQKFGDRAYKAWALVRLSQRFPKADFTDISKIVTDLTK
BSA        KIETMREKVLTSSARQRLRCASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTK
           *:: :. .   *:*::**:*:*:*:**:** ***::.****:****:*.:::*:******

HSA        VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPA
ESA        VHKECCHGDLLECADDRADLAKYICEHQDSISGKLKACCDKPLLQKSHCIAEVKEDDLPS
LSA        VHKECCHGDLLECADDRADLAKYMCEHQETISSHLKECCDKPILEKAHCIYGLHNDETPA
RSA        VHKECCHGDLLECADDRADLAKYMCEHQETISSHLKECCDKPILEKAHCIYGLHNDETPA
BSA        VHKECCHGDLLECADDRADLAKYICDNQDTISSKLKECCDKPLLEKSHCIAEVEKDAIPE
           **.********************:*:.*::**.:** **:**:*:*:***  :.:*  *

HSA        DLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKC
ESA        DLPALAADFAEDKEICKHYKDAKDVFLGTFLYEYSRRHPDYSVSLLLRIAKTYEATLEKC
LSA        GLPAVAEEFVEDKDVCKNYEEAKDLFLGKFLYEYSRRHPDYSVVLLLRLGKAYEATLKKC
RSA        GLPAVAEEFVEDKDVCKNYEEAKDLFLGKFLYEYSRRHPDYSVVLLLRLGKAYEATLKKC
BSA        NLPPLTADFAEDKDVCKNYQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEATLEEC
            ** :: :*.*.*::**.* :*** *** *****:****:*:* :***:.* **:**::*

HSA        CAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST
ESA        CAEADPPACYRTVFDQFTPLVEEPKSLVKKNCDLFEEVGEYDFQNALIVRYTKKAPQVST
LSA        CATDDPHACYAKVLDEFQPLVDEPKNLVKQNCELYEQLGDYNFQNALLVRYTKKVPQVST
RSA        CATDDPHACYAKVLDEFQPLVDEPKNLVKQNCELYEQLGDYNFQNALLVRYTKKVPQVST
BSA        CAKDDPHACYSTVFDKLKHLVDEPQNLIKQNCDQFEKLGEYGFQNALIVRYTRKVPQVST
           **  **  ** .*:*::  **:**:.*:*:**: :*::*:* *****:****:*.*****

HSA        PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTES
ESA        PTLVEIGRTLGKVGSRCCKLPESERLPCSENHLALALNRLCVLHEKTPVSEKITKCCTDS
LSA        PTLVEISRSLGKVGSKCCKHPEAERLPCVEDYLSVVLNRLCVLHEKTPVSEKVTKCCSES
RSA        PTLVEISRSLGKVGSKCCKHPEAERLPCVEDYLSVVLNRLCVLHEKTPVSEKVTKCCSES
BSA        PTLVEVSRSLGKVGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTES
           *****:.*.*****::**. **::*:** *::*:: **:***********:::****::*

HSA        LVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKAT
ESA        LAERRPCFSALELDEGYVPKEFKAETFTFHADICTLPEDEKQIKKQSALAELVKHKPKAT
LSA        LVDRRPCFSALGPDETYVPKEFNAETFTFHADICTLPETERKIKKQTALVELVKHKPHAT
RSA        LVDRRPCFSALGPDETYVPKEFNAETFTFHADICTLPETERKIKKQTALVELVKHKPHAT
BSA        LVNRRPCFSALTPDETYVPKAFDEKLFTFHADICTLPDTEKQIKKQTALVELLKHKPKAT
           *.:********  ** **** *. : ********** : *::****:**.**:****:**

HSA        KEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
ESA        KEQLKTVLGNFSAFVAKCCGREDKEACFAEEGPKLVASSQLALA-
LSA        NDQLKTVVGEFTALLDKCCSAEDKEACFAVEGPKLVESSKATLG-
RSA        NDQLKTVVGEFTALLDKCCSAEDKEACFAVEGPKLVESSKATLG-
BSA        EEQLKTVMENFVAFVDKCCAADDKEACFAVEGPKLVVSTQTALA-
           ::***:*: :* *:: ***  :***:*** ** *** ::: :*.
```

From the “Result Summary” tab, the percent identity matrix can be obtained:

	HSA	ESA	LSA	RSA	BSA
HSA	100.00	76.33	74.32	74.32	75.64

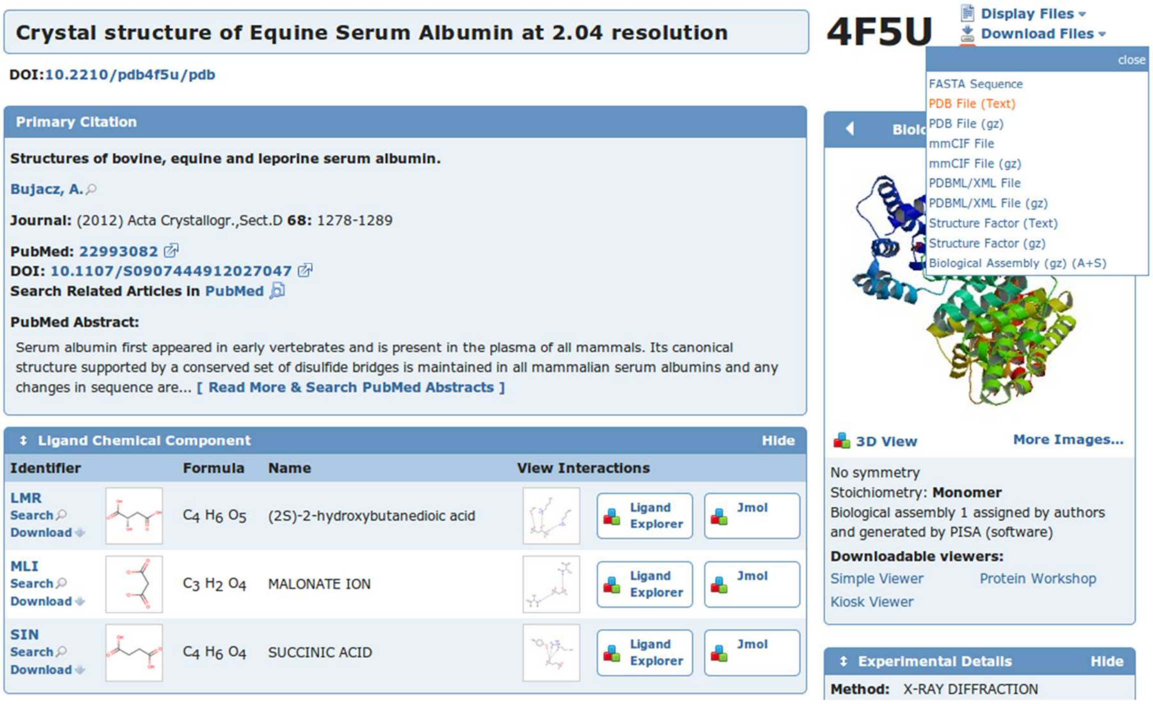
	HSA	ESA	LSA	RSA	BSA
ESA	76.33	100.00	71.01	70.67	73.93
LSA	74.32	71.01	100.00	99.49	71.53
RSA	74.32	70.67	99.49	100.00	71.36
BSA	75.64	73.93	71.53	71.36	100.00

From the percent identity matrix conclusion that equine serum albumin is the most similar to the human serum albumin can be made. Therefore, ESA model should be used for building a homology structure.

PART IV: PREPARING TEMPLATE FOR HOMOLOGY MODELING

After decision to use ESA as template for homology modeling, PDB file from the RCSB Protein Data Bank should be obtained. Since there are four PDB entries (3V08, 4F5T, 4F5U, 4J2V), structure with PDB ID: 4F5U should be selected and textual PDB file downloaded.

Among four structures, 4F5U has the highest resolution (2.04 Å). All other structures have lower resolutions: 4J2V (2.12 Å), 4F5T (2.32 Å) and 3V08 (2.45 Å). As the rule of thumb, structures with higher resolution (lower number of angstroms) are usually better for modeling.



Crystal structure of Equine Serum Albumin at 2.04 resolution

DOI:10.2210/pdb4f5u/pdb

Primary Citation

Structures of bovine, equine and leporine serum albumin.

Bujacz, A.

Journal: (2012) Acta Crystallogr.,Sect.D **68:** 1278-1289

PubMed: 22993082
DOI: 10.1107/S0907444912027047
Search Related Articles in PubMed

PubMed Abstract:

Serum albumin first appeared in early vertebrates and is present in the plasma of all mammals. Its canonical structure supported by a conserved set of disulfide bridges is maintained in all mammalian serum albumins and any changes in sequence are... [Read More & Search PubMed Abstracts]

Ligand Chemical Component Hide

Identifier		Formula	Name	View Interactions		
LMR Search Download		$C_4H_6O_5$	(2S)-2-hydroxybutanedioic acid		Ligand Explorer	Jmol
MLI Search Download		$C_3H_2O_4$	MALONATE ION		Ligand Explorer	Jmol
SIN Search Download		$C_4H_6O_4$	SUCCINIC ACID		Ligand Explorer	Jmol

4F5U

No symmetry
Stoichiometry: **Monomer**
Biological assembly 1 assigned by authors and generated by PISA (software)

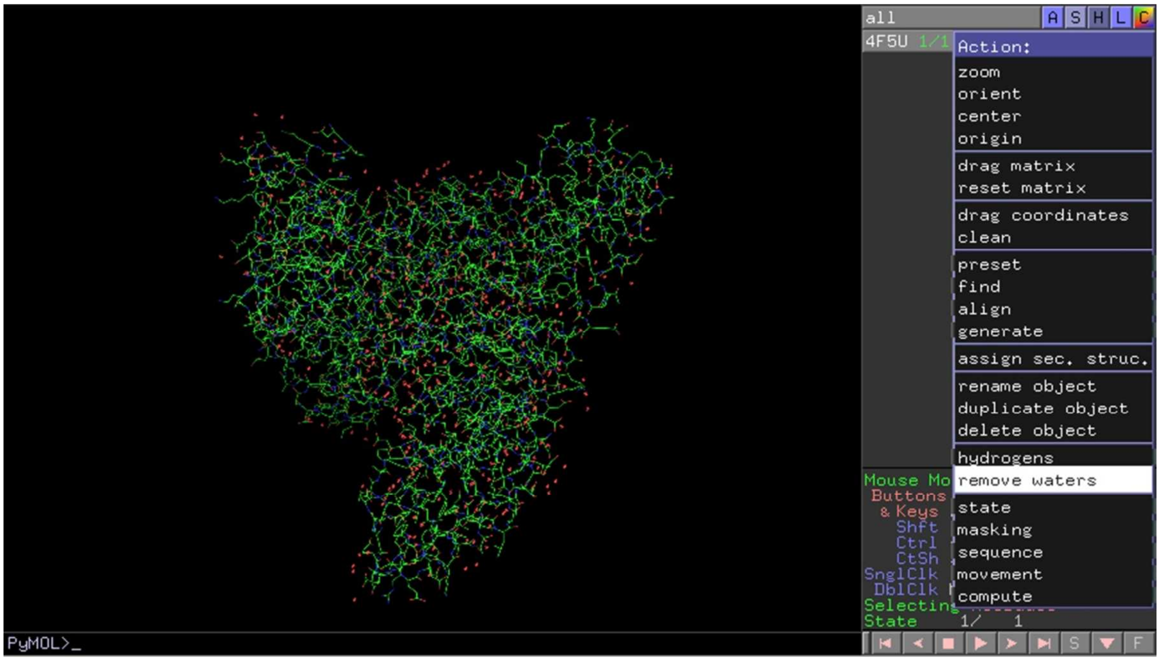
Downloadable viewers:
Simple Viewer Protein Workshop
Kiosk Viewer

Experimental Details Hide

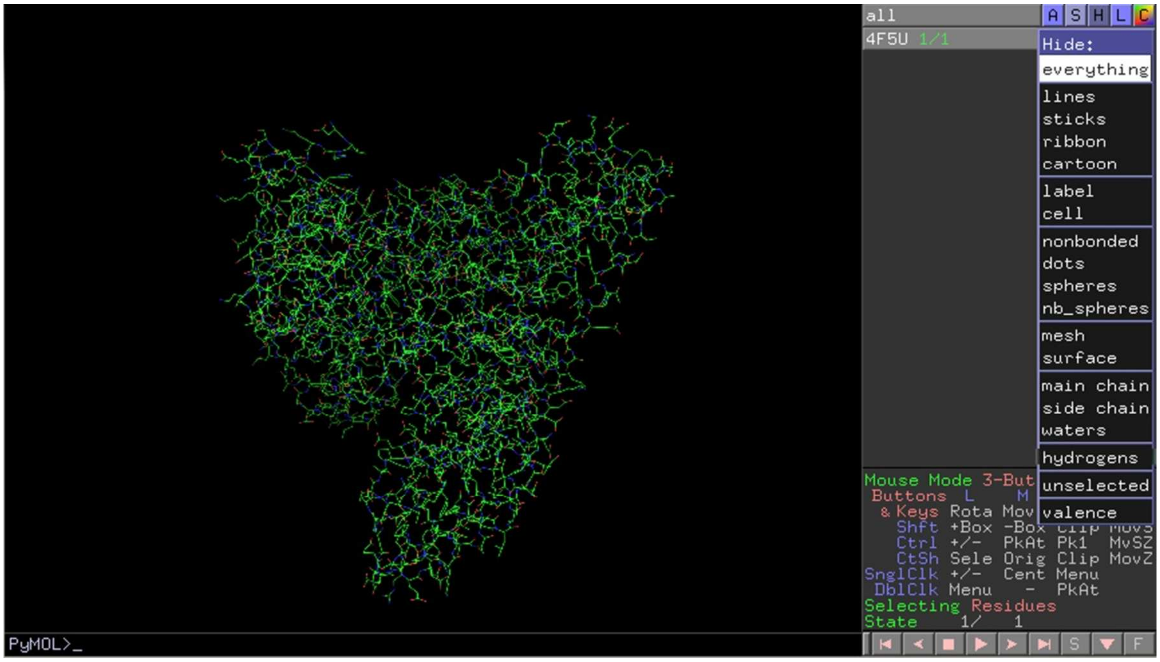
Method: X-RAY DIFFRACTION

Equine serum albumin 4F5U was co-crystallized with 8 ligands: one molecule of (2*S*)-2-hydroxybutanedioic acid (LMR), six malonate ions (MLI) and one molecule of succinic acid (SIN), and with 345 water molecules. To build a homology model, all ligands and water molecules should be removed using PyMOL (or your preferred molecular editor).

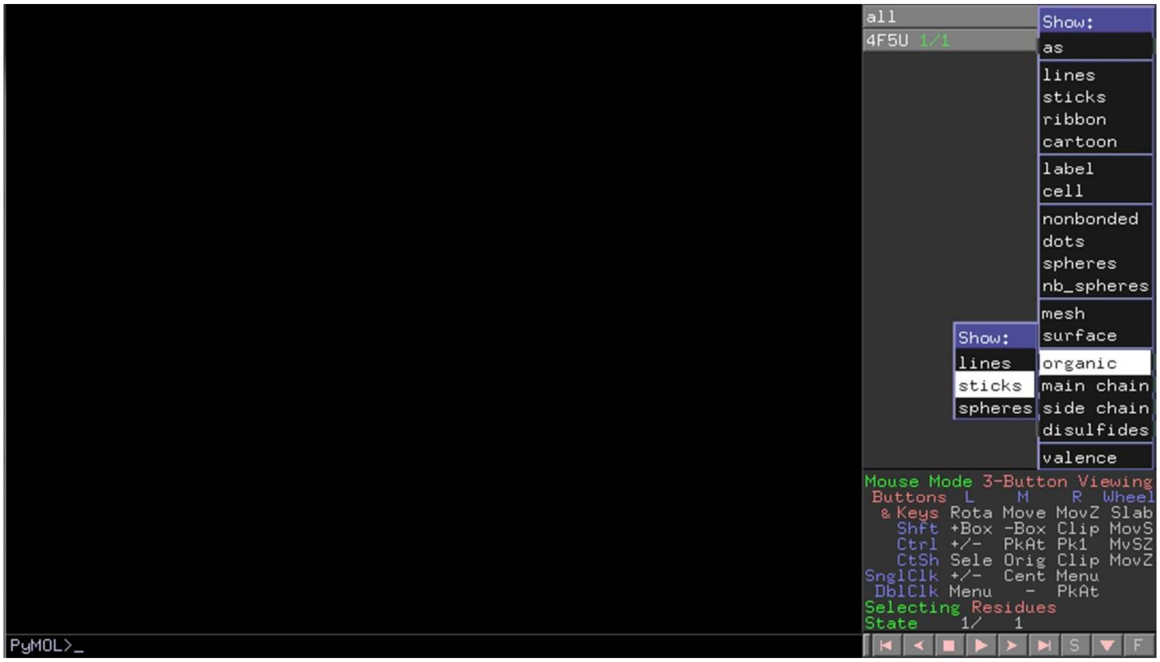
In PyMOL, click on the A (action) button of the 4F5U and choose "remove water" option.



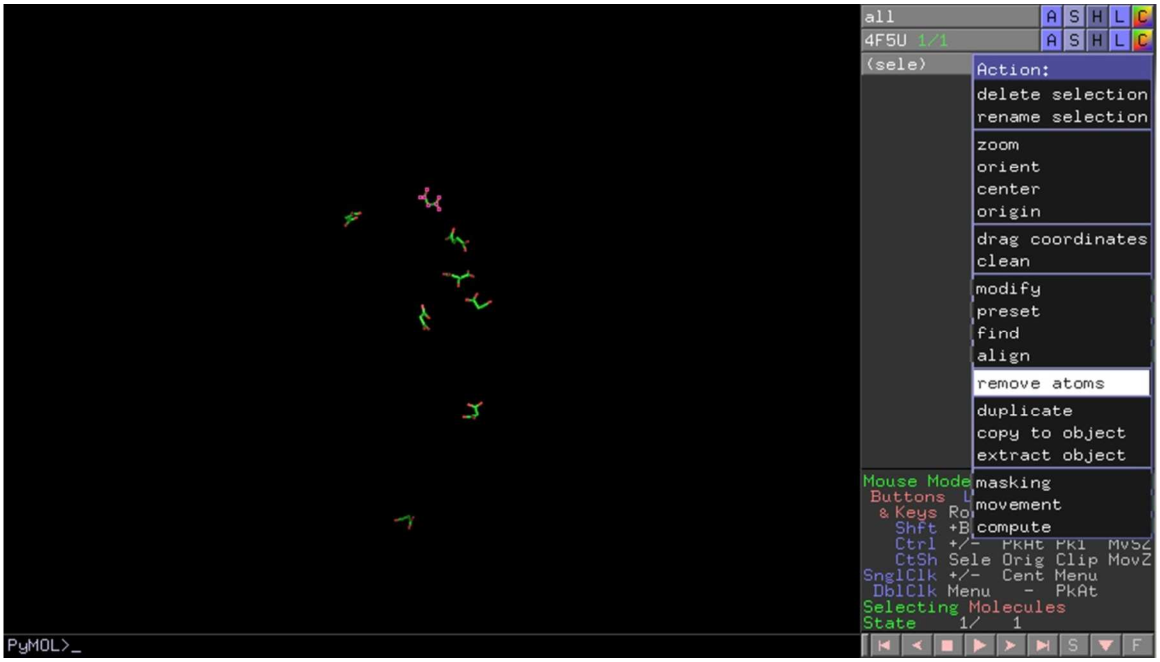
In order to visualize ligands, first hide everything:



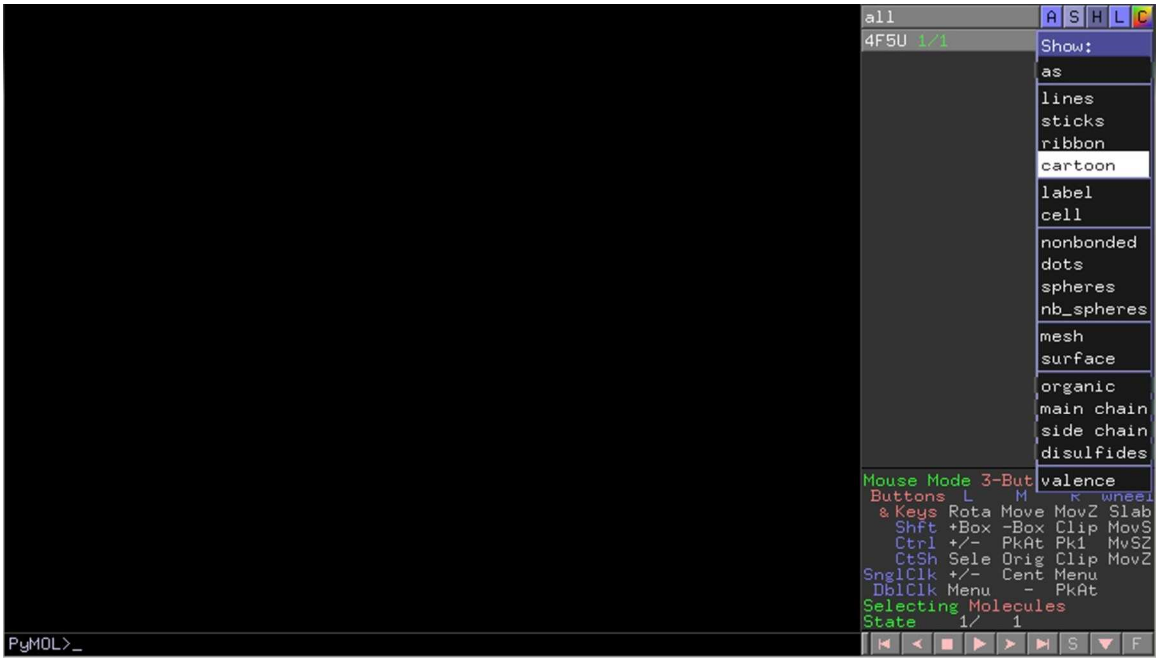
and then show organic molecules as sticks:



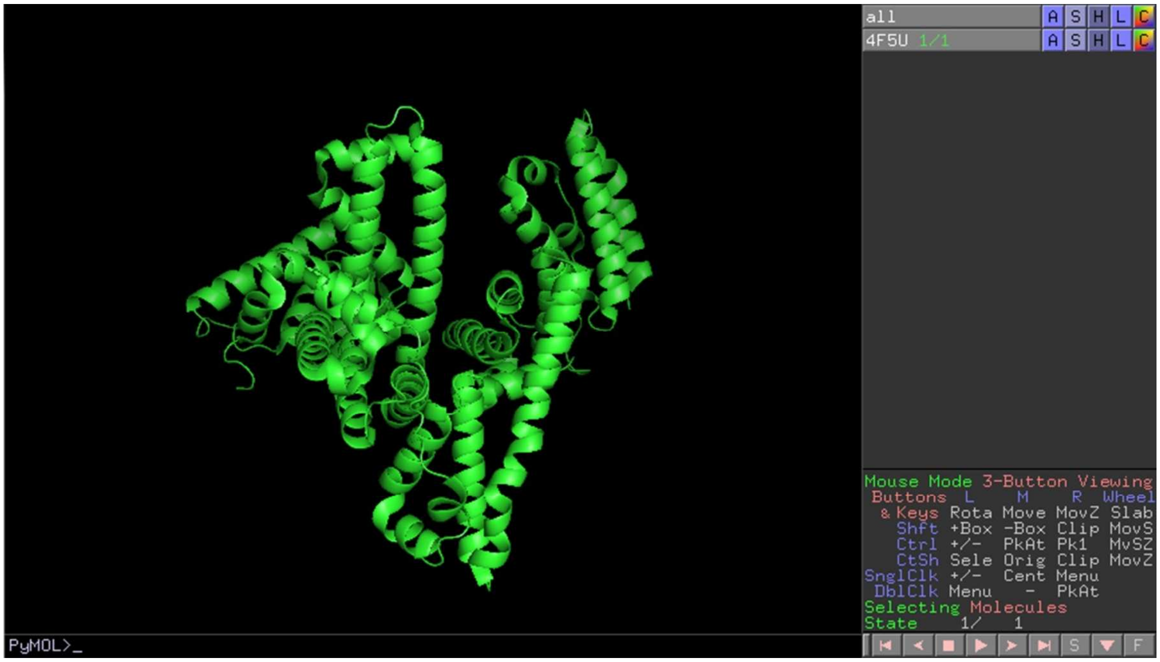
Finally, delete ligands one by one. In the bottom right part of the menu click on the “Selecting” command until “Molecules” have been chosen.



Choose one of the ligand molecules, and (sele) section will appear. From the action button choose “remove atoms” command. After deleting 8 ligands, show cartoon of the 4F5U:



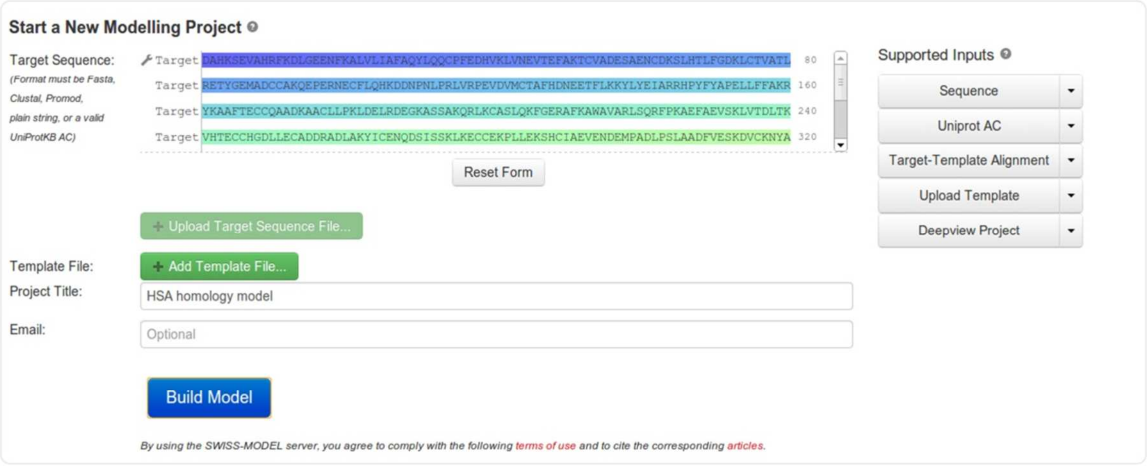
and from the File menu choose “Save molecule” option. Newly saved PDB file will be used as input for homology modeling.



PART V: BUILDING A HOMOLOGY MODEL

For building a homology model SWISS-MODEL web server will be used. It is available at ExPASy (http://swissmodel.expasy.org/interactive). Models are built based on the target-template alignment using Promod-II. Coordinates which are conserved between the target and the template are copied from the template to the model. Insertions and deletions are remodeled using a fragment library. Side chains are then rebuilt. Finally, the geometry of the resulting model is regularized by using a force field.

To build a model, first click on the *Upload Template* button on the right. Paste HSA sequence to the "Target Sequence" section. To add template file, click the *Add Template File...* button and choose previously prepared PDB structure. After "Template Uploaded ✓" sign appears, provide a project title and click the *Build Model* button.



Start a New Modelling Project

Target Sequence: (Format must be Fasta, Clustal, Promod, plain string, or a valid UniProtKB AC)

```
Target DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATL  80
Target RETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKR 160
Target YKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK 240
Target VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYA 320
```

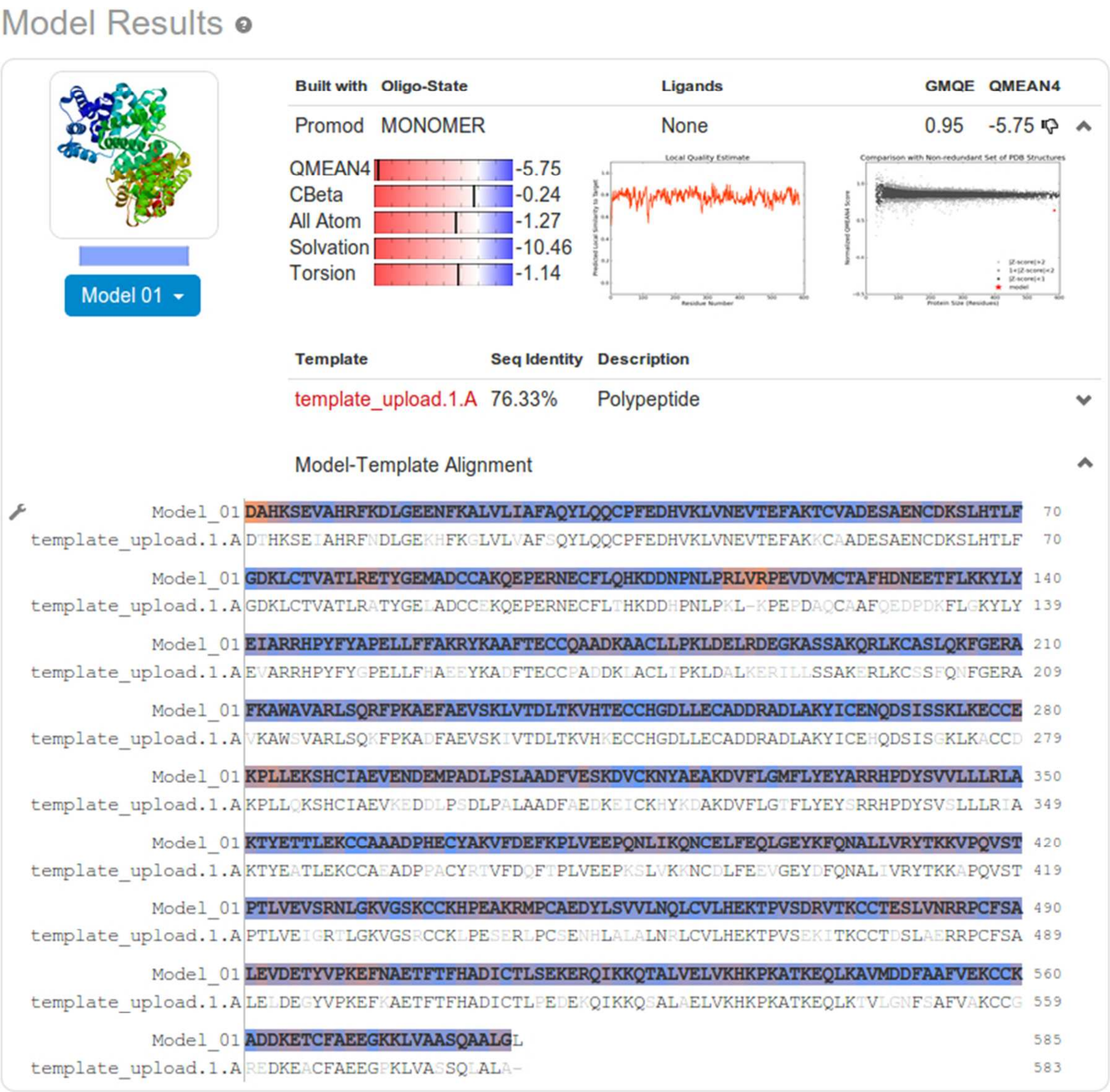
Template File:

Project Title: HSA homology model

Email: Optional

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

The model result page appears with some model analysis and with model-template alignment.



Model Results

Built with	Oligo-State	Ligands	GMQE	QMEAN4
Promod	MONOMER	None	0.95	-5.75

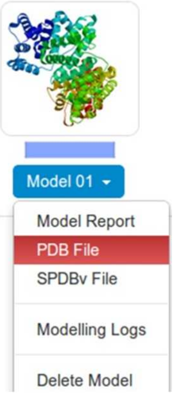
QMEAN4	-5.75
CBeta	-0.24
All Atom	-1.27
Solvation	-10.46
Torsion	-1.14

Template	Seq Identity	Description
template_upload.1.A	76.33%	Polypeptide

Model-Template Alignment

```
          Model_01 DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLF  70
template_upload.1.A DTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPFEDHVKLVNEVTEFAKKCAADESAENCDKSLHTLF  70

          Model_01 GDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLY 140
template_upload.1.A GDKLCTVATLRATYGELADCCEKQEPERNECFLTHKDDHPNLPKL-KPEPDAQCAAFQEDPDKFLGKYLY 139

          Model_01 EIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERA 210
template_upload.1.A EVARRHPYFYGPELLFHAEEYKADFTECCPADDKLACLIPKLDALKERILLSSAKERLKCSSFQNFGERA 209

          Model_01 FKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCE 280
template_upload.1.A VKAWSVARLSQKFPKADFAEVSKIVTDLTKVHKECCHGDLLECADDRADLAKYICEHQDSISGKLKACCD 279

          Model_01 KPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLA 350
template_upload.1.A KPLLQKSHCIAEVKEDDLPSDLPALAADFAEDKEICKHYKDAKDVFLGTFLYEYSRRHPDYSVSLLLRIA 349

          Model_01 KTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST 420
template_upload.1.A KTYEATLEKCCAEADPPACYRTVFDQFTPLVEEPKSLVKKNCDLFEEVGEYDFQNALIVRYTKKAPQVST 419

          Model_01 PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSA 490
template_upload.1.A PTLVEIGRTLGKVGSRCCKLPESERLPCSENHLALALNRLCVLHEKTPVSEKITKCCTDSLAERRPCFSA 489

          Model_01 LEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCK 560
template_upload.1.A LELDEGYVPKEFKAETFTFHADICTLPEDEKQIKKQSALAELVKHKPKATKEQLKTVLGNFSAFVAKCCG 559

          Model_01 ADDKETCFAEEGKKLVAASQAALGL                                              585
template_upload.1.A REDKEACFAEEGPKLVASSQLALA-                                              583
```

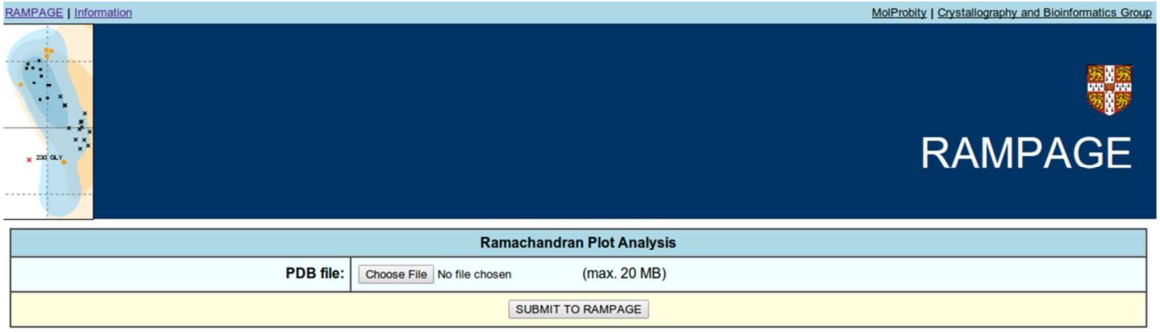
To download the PDB file, choose “PDB File” option from “Model 01” menu.



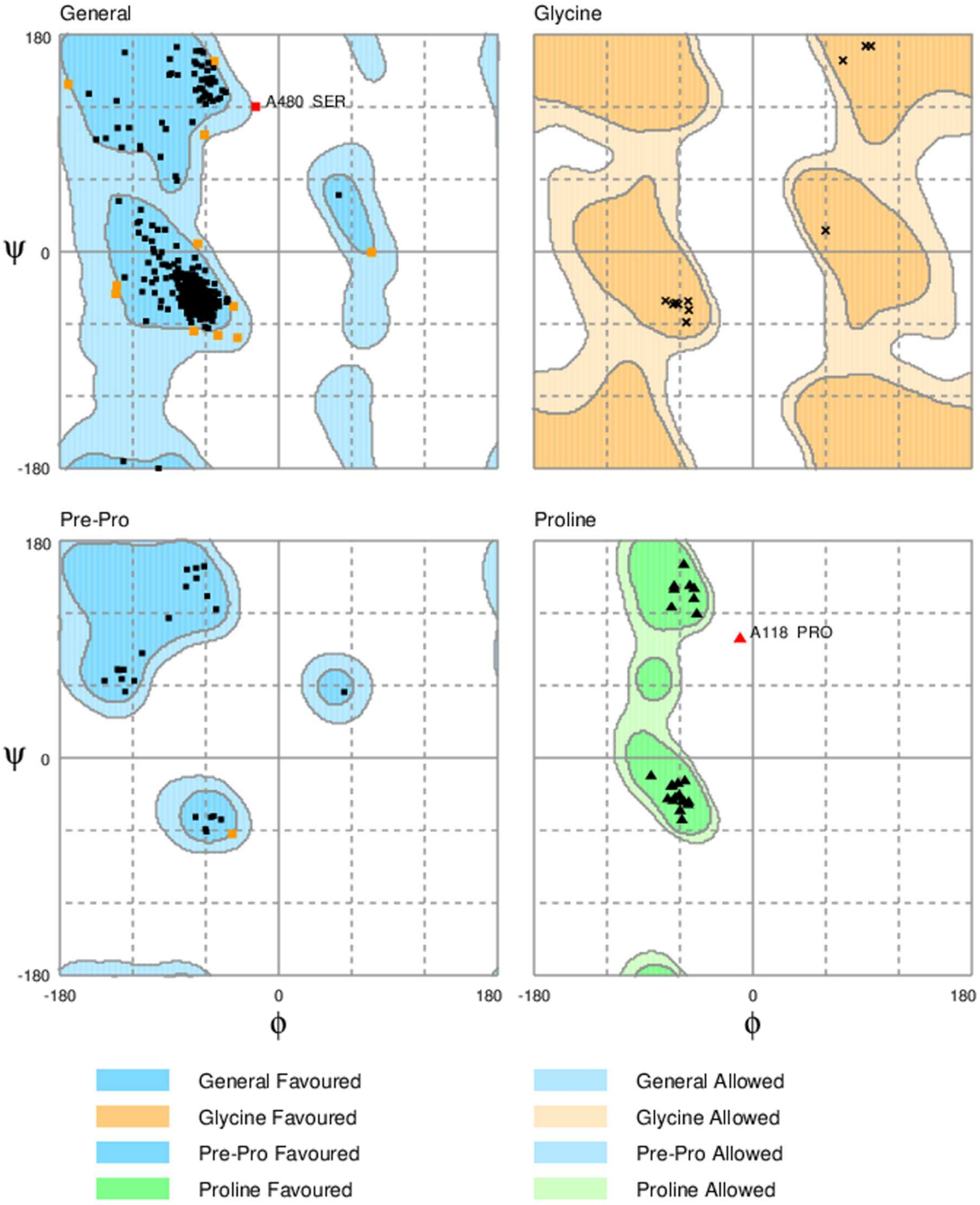
PART VI: ANALYZING BUILT HOMOLOGY MODEL

After a homology model is built, its quality should be tested by examining protein's geometry. One of the easiest tools to visually analyze torsion angles is Ramachandran plot. Although many molecular modeling software applications can prepare this type of plot, an on-line tool RAMPAGE available at http://mordred.bioc.cam.ac.uk/~rapper/rampage.php will be used.

To perform analysis press *Choose File* button, navigate to HSA_model.pdb file and finally press the *SUBMIT TO RAMPAGE* button. Results of visual and numerical analysis appear.



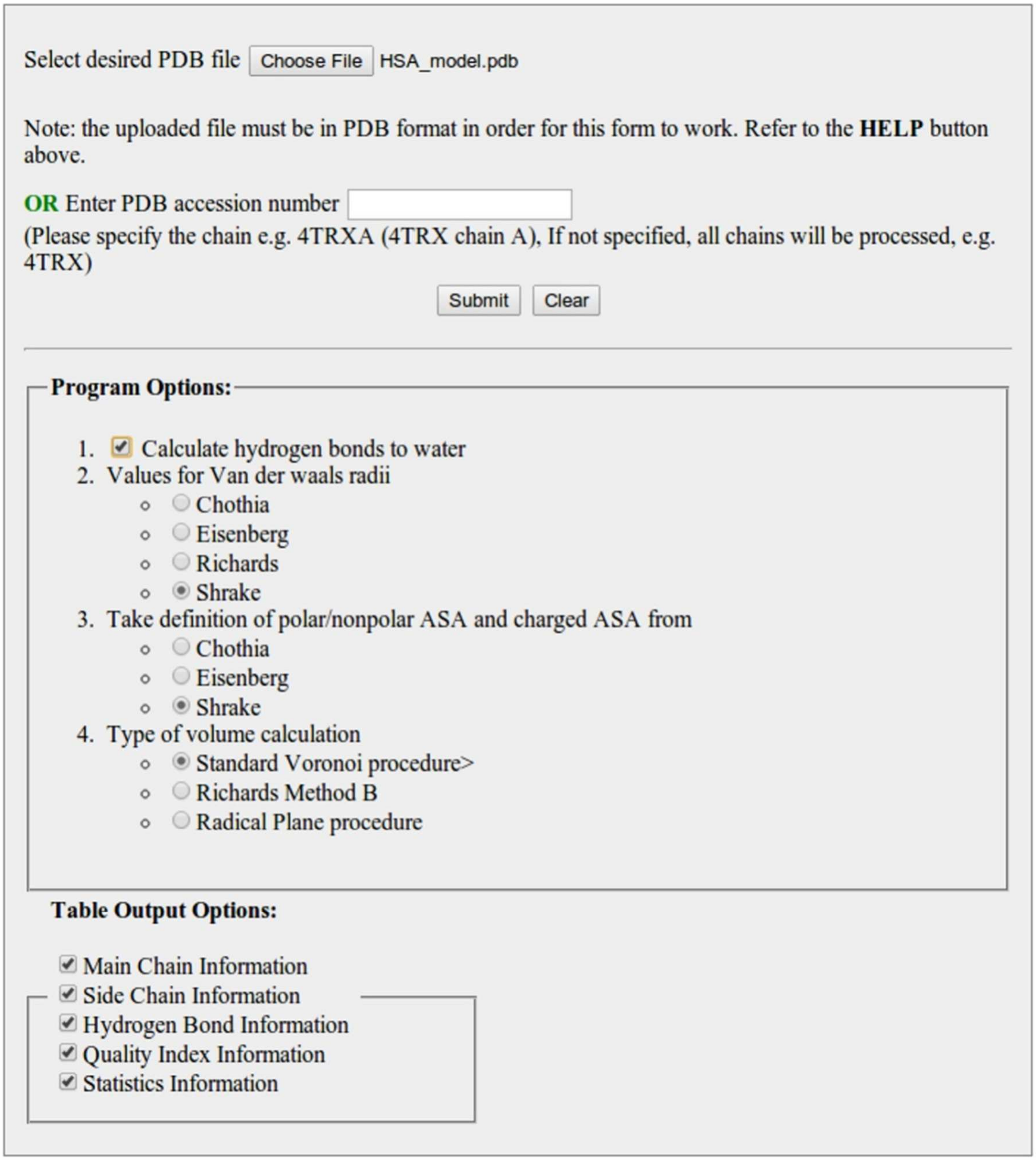
```
Number of residues in favored region    (~98.0% expected):  568 (97.6%)
Number of residues in allowed region    ( ~2.0% expected):   12  (2.1%)
Number of residues in outlier region                     :    2  (0.3%)


Residue [A   4  LYS] ( -37.17, -45.42) in Allowed region
Residue [A  61  ASN] (  76.30,  -0.53) in Allowed region
Residue [A  65  SER] ( -53.15, 157.99) in Allowed region
Residue [A 150  TYR] ( -61.09,  97.12) in Allowed region
Residue [A 151  ALA] ( -38.14, -62.95) in Allowed region
Residue [A 272  SER] (-173.05, 138.77) in Allowed region
Residue [A 283  LEU] ( -33.95, -71.40) in Allowed region
Residue [A 310  VAL] (-133.87, -34.74) in Allowed region
Residue [A 320  ALA] ( -69.83, -65.88) in Allowed region
Residue [A 323  LYS] ( -50.04, -69.40) in Allowed region
Residue [A 469  VAL] (-133.23, -28.11) in Allowed region
Residue [A 495  GLU] ( -66.44,   6.60) in Allowed region
Residue [A 118  PRO] ( -10.55, 100.06) in Outlier region
Residue [A 480  SER] ( -18.84, 120.34) in Outlier region
```



Besides Ramachandran plot inspection, other properties of the HSA model have to be analyzed using VADAR (Volume, Area, Dihedral Angle Reporter) software (http://vadar.wishartlab.com/).

VADAR compares results calculated for the analyzed protein with expected values extracted from highly refined X-ray and NMR protein structures.

To calculate protein properties, click on the *Choose File* button and navigate to HSA_model.pdb file. Also, check "Calculate hydrogen bonds to water" option and finally click the *Submit* button.



Select desired PDB file HSA_model.pdb

Note: the uploaded file must be in PDB format in order for this form to work. Refer to the **HELP** button above.

OR Enter PDB accession number

(Please specify the chain e.g. 4TRXA (4TRX chain A), If not specified, all chains will be processed, e.g. 4TRX)

Program Options:

1. ☑ Calculate hydrogen bonds to water
2. Values for Van der waals radii
   - ○ Chothia
   - ○ Eisenberg
   - ○ Richards
   - ● Shrake
3. Take definition of polar/nonpolar ASA and charged ASA from
   - ○ Chothia
   - ○ Eisenberg
   - ● Shrake
4. Type of volume calculation
   - ● Standard Voronoi procedure>
   - ○ Richards Method B
   - ○ Radical Plane procedure

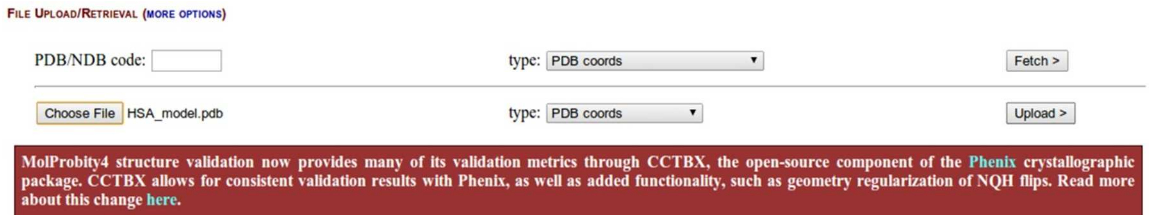
Table Output Options:

☑ Main Chain Information
☑ Side Chain Information
☑ Hydrogen Bond Information
☑ Quality Index Information
☑ Statistics Information

We will partially examine "Statistics" output file.

```
                     **********************************
                     *           VADAR STATS          *
                     **********************************
                    ** Using atomic radii from Shrake **


|------------------------------|---------------------|---------------------|
| Statistic                    | Observed            | Expected            |
|------------------------------|---------------------|---------------------|
| # Helix                      |    433   ( 74%)     |         -           |
| # Beta                       |      6   (  1%)     |         -           |
| # Coil                       |    145   ( 24%)     |         -           |
|------------------------------|---------------------|---------------------|
| # Turn                       |    136   ( 23%)     |         -           |
|------------------------------|---------------------|---------------------|

                          HYDROGEN BONDS (hbonds)

|------------------------------|---------------------|---------------------|
| Statistic                    | Observed            | Expected            |
|------------------------------|---------------------|---------------------|
|  Meanhbond distance          |      2.1 sd=0.4     |      2.2 sd=0.4     |
|  Meanhbond energy            |     -2.1 sd=1.3     |     -2.0 sd=0.8     |
|  # res with hbonds           |    529   ( 90%)     |    438   ( 75%)     |
|------------------------------|---------------------|---------------------|

                              DIHEDRAL ANGLES

|------------------------------|---------------------|---------------------|
| Statistic                    | Observed            | Expected            |
|------------------------------|---------------------|---------------------|
|  Mean Helix Phi              |    -66.8 sd=9.5     |    -65.3 sd=11.9    |
|  Mean Helix Psi              |    -38.8 sd=12.8    |    -39.4 sd=25.5    |
|  # res with Gauche+ Chi      |    238   ( 48%)     |    267   ( 55%)     |
|  # res with Gauche- Chi      |     65   ( 13%)     |     97   ( 20%)     |
|  # res with Trans Chi        |    183   ( 37%)     |    121   ( 25%)     |
|  Mean Chi Gauche+            |    -67.3 sd=9.8     |    -66.7 sd=15.0    |
|  Mean Chi Gauche-            |     65.4 sd=6.4     |     64.1 sd=15.7    |
|  Mean Chi Trans              |    172.6 sd=6.4     |    168.6 sd=16.8    |
|  Std. dev of chi pooled      |         8.08        |         15.70       |
|  Mean Omega (|omega|>90)     |    179.0 sd=5.0     |    180.0 sd=5.8     |
|  # res with |omega|<90       |      2   (  0%)     |         -           |
|------------------------------|---------------------|---------------------|
                         ACCESSIBLE SURFACE AREA (ASA)

|------------------------------|---------------------|---------------------|
| Statistic                    | Observed            | Expected            |
|------------------------------|---------------------|---------------------|
|  Total ASA                   |   23180.7 Angs**2   |  20606.5 Angs**2    |
|  ASA of backbone             |    1925.3 Angs**2   |         -           |
|  ASA of sidechains           |   21255.4 Angs**2   |         -           |
|  ASA of C                    |   14536.7 Angs**2   |         -           |
|  ASA of N                    |     980.9 Angs**2   |         -           |
|  ASA of N+                   |    1504.4 Angs**2   |         -           |
|  ASA of O                    |    3998.5 Angs**2   |         -           |
|  ASA of O-                   |    2108.9 Angs**2   |         -           |
|  ASA of S                    |      51.3 Angs**2   |         -           |
|  Exposed nonpolar ASA        |   14144.0 Angs**2   |  14140.2 Angs**2    |
|  Exposed polar ASA           |    3092.9 Angs**2   |   4636.1 Angs**2    |
|  Exposed charged ASA         |    5943.8 Angs**2   |   4404.3 Angs**2    |
|  Side exposed nonpolar ASA   |   14177.2 Angs**2   |         -           |
|  Side exposed polar ASA      |    1225.3 Angs**2   |         -           |
|  Side exposed charged ASA    |    5852.9 Angs**2   |         -           |
|  Fraction nonpolar ASA       |        0.61         |     0.61 sd=0.03    |
|  Fraction polar ASA          |        0.13         |     0.20 sd=0.05    |
|  Fraction charged ASA        |        0.26         |     0.19 sd=0.05    |
|  Mean residue ASA            |     39.7 sd=40.6    |         -           |
|  Meanfrac ASA                |      0.2 sd=0.2     |         -           |
|  % side ASA hydrophobic      |       22.23         |         -           |
|------------------------------|---------------------|---------------------|

                                    VOLUME

|------------------------------|---------------------|---------------------|
| Statistic                    | Observed            | Expected            |
|------------------------------|---------------------|---------------------|
|  Total volume (packing)      |   80043.2 Angs**3   |  80779.4 Angs**3    |
|  Mean residue volume         |    137.1 sd=46.7    |   125.0 sd=40.0     |
|  Meanfrac volume             |      1.0 sd=0.3     |     1.0 sd=0.1      |
|  Molecular weight            |    66361.24         |         -           |
|------------------------------|---------------------|---------------------|


                              ***************
                              *  END VADAR  *
                              ***************
```

Some further analysis of the HSA model can be performed at the MolProbity web server (http://molprobity.biochem.duke.edu/). At the main page, click at the *Choose File* button and navigate to the HSA_model PDB file. Click the *Upload >* button to start analyzing model.

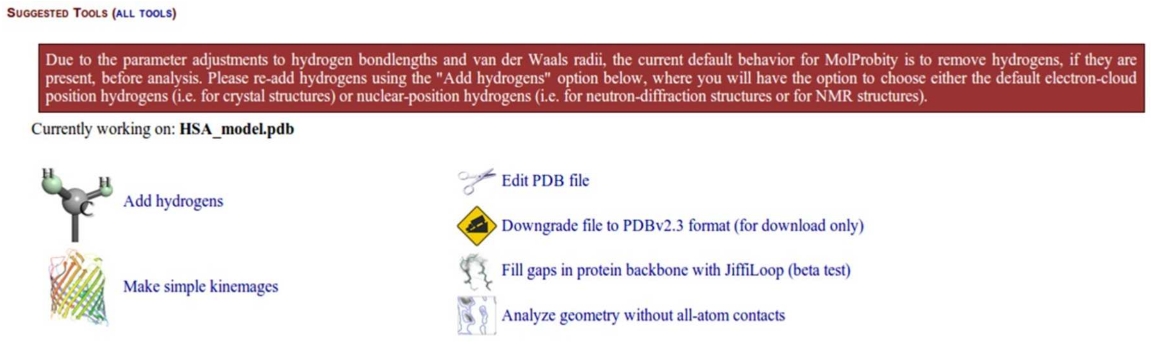


File Upload/Retrieval (more options)

PDB/NDB code:	type: PDB coords	Fetch >
Choose File HSA_model.pdb	type: PDB coords	Upload >

MolProbity4 structure validation now provides many of its validation metrics through CCTBX, the open-source component of the Phenix crystallographic package. CCTBX allows for consistent validation results with Phenix, as well as added functionality, such as geometry regularization of NQH flips. Read more about this change here.

From the tool panel choose *Analyze geometry without all-atom contacts* option,



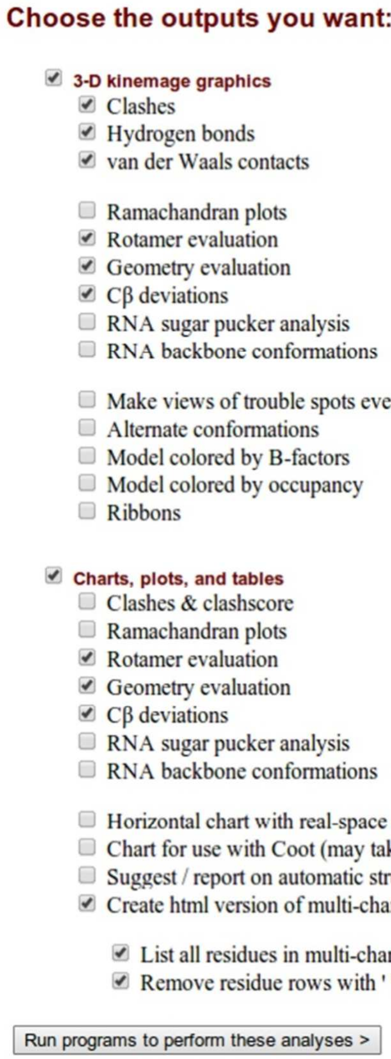
SUGGESTED TOOLS (ALL TOOLS)

Due to the parameter adjustments to hydrogen bondlengths and van der Waals radii, the current default behavior for MolProbity is to remove hydrogens, if they are present, before analysis. Please re-add hydrogens using the "Add hydrogens" option below, where you will have the option to choose either the default electron-cloud position hydrogens (i.e. for crystal structures) or nuclear-position hydrogens (i.e. for neutron-diffraction structures or for NMR structures).

Currently working on: **HSA_model.pdb**

Add hydrogens
Make simple kinemages
Edit PDB file
Downgrade file to PDBv2.3 format (for download only)
Fill gaps in protein backbone with JiffiLoop (beta test)
Analyze geometry without all-atom contacts

and fill the form for the outputs you would like to get.



Choose the outputs you want:

☑ **3-D kinemage graphics**
- ☑ Clashes
- ☑ Hydrogen bonds
- ☑ van der Waals contacts

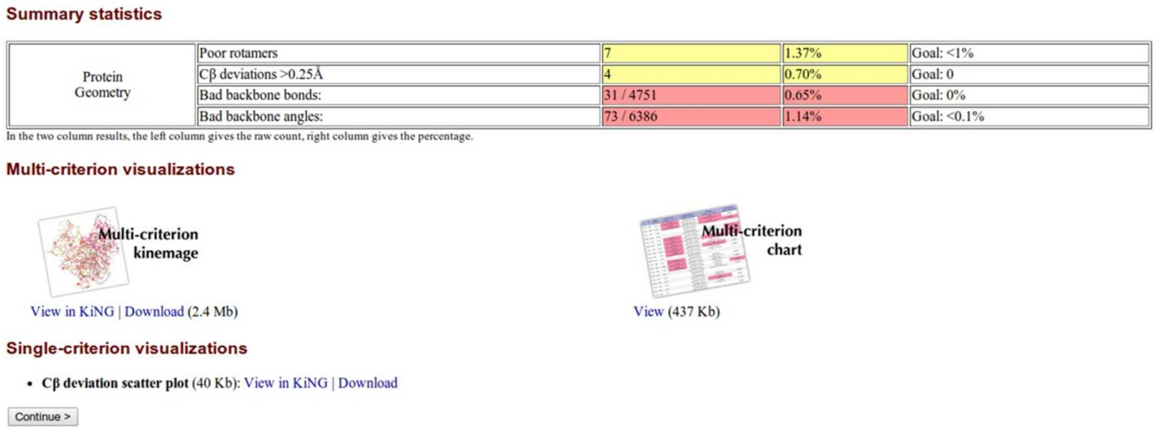
- □ Ramachandran plots
- ☑ Rotamer evaluation
- ☑ Geometry evaluation
- ☑ Cβ deviations
- □ RNA sugar pucker analysis
- □ RNA backbone conformations

- □ Make views of trouble spots even if it takes longer
- □ Alternate conformations
- □ Model colored by B-factors
- □ Model colored by occupancy
- □ Ribbons

☑ **Charts, plots, and tables**
- □ Clashes & clashscore
- □ Ramachandran plots
- ☑ Rotamer evaluation
- ☑ Geometry evaluation
- ☑ Cβ deviations
- □ RNA sugar pucker analysis
- □ RNA backbone conformations

- □ Horizontal chart with real-space correlation data
- □ Chart for use with Coot (may take a long time, but should take less than 1 hour)
- □ Suggest / report on automatic structure fix-ups
- ☑ Create html version of multi-chart
  - ☑ List all residues in multi-chart, not just outliers
  - ☑ Remove residue rows with ' ' altloc when other alternate(s) present

At the result page, take care about summary statistics. To analyze multi-criterion kinemage, click on the *View in KING* button.



Summary statistics

Protein Geometry	Poor rotamers	7	1.37%	Goal: <1%
	Cβ deviations >0.25Å	4	0.70%	Goal: 0
	Bad backbone bonds:	31 / 4751	0.65%	Goal: 0%
	Bad backbone angles:	73 / 6386	1.14%	Goal: <0.1%

In the two column results, the left column gives the raw count, right column gives the percentage.

Multi-criterion visualizations

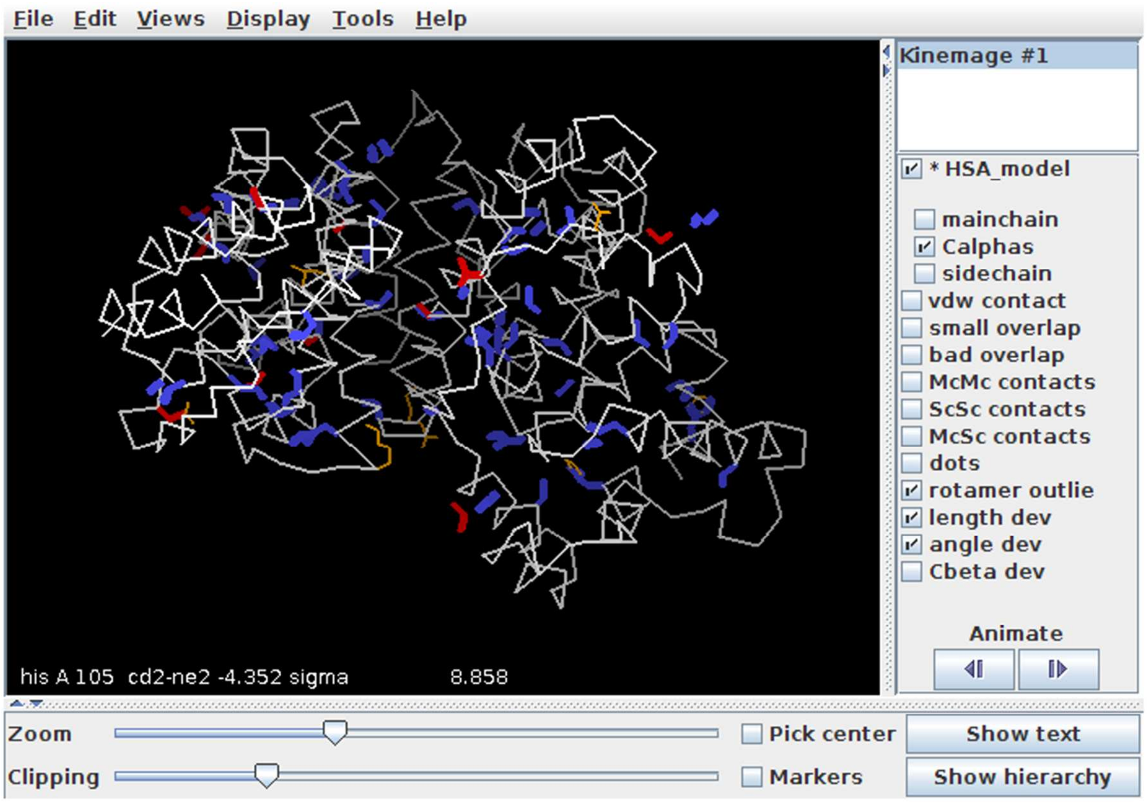
View in KiNG | Download (2.4 Mb)

View (437 Kb)

Single-criterion visualizations

- **Cβ deviation scatter plot** (40 Kb): View in KiNG | Download

Explore rotamer outliers as well as bond length and angle deviations. To identify amino acid residue click on the line and residue information will appear in the bottom left part of the page.



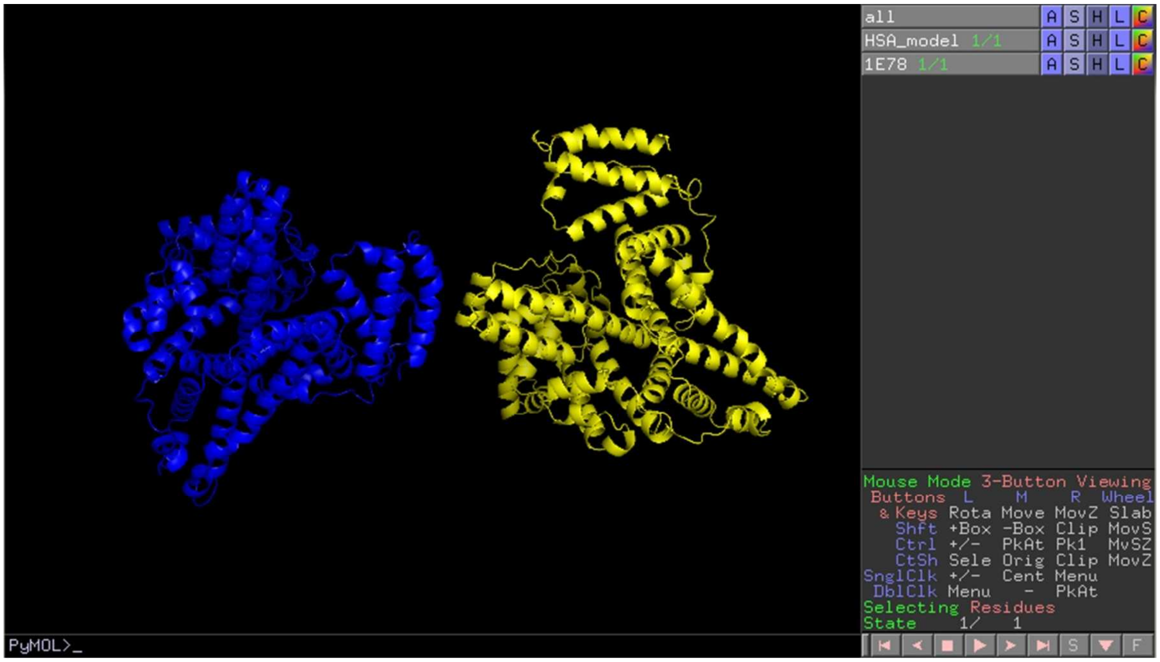
☑ * HSA_model
□ mainchain
☑ Calphas
□ sidechain
□ vdw contact
□ small overlap
□ bad overlap
□ McMc contacts
□ ScSc contacts
□ McSc contacts
□ dots
☑ rotamer outlie
☑ length dev
☑ angle dev
□ Cbeta dev

□ Pick center
□ Markers

```
his A 105  cd2-ne2 -4.352 sigma          8.858
```

PART VII: BENCHMARKING HOMOLOGY MODEL VS. PDB DEPOSITED STRUCTURES

The good way to check how similar two structures are is to align them and to calculate the root mean square deviation (RMSD) between atoms coordinates. To calculate RMSD in PyMOL, open HSA_model and 1E78 (human serum albumin without co-crystallized ligands) structures. Show them as cartoons only, and color HSA_model to blue (using the C button), and benchmark 1E78 structure to yellow.

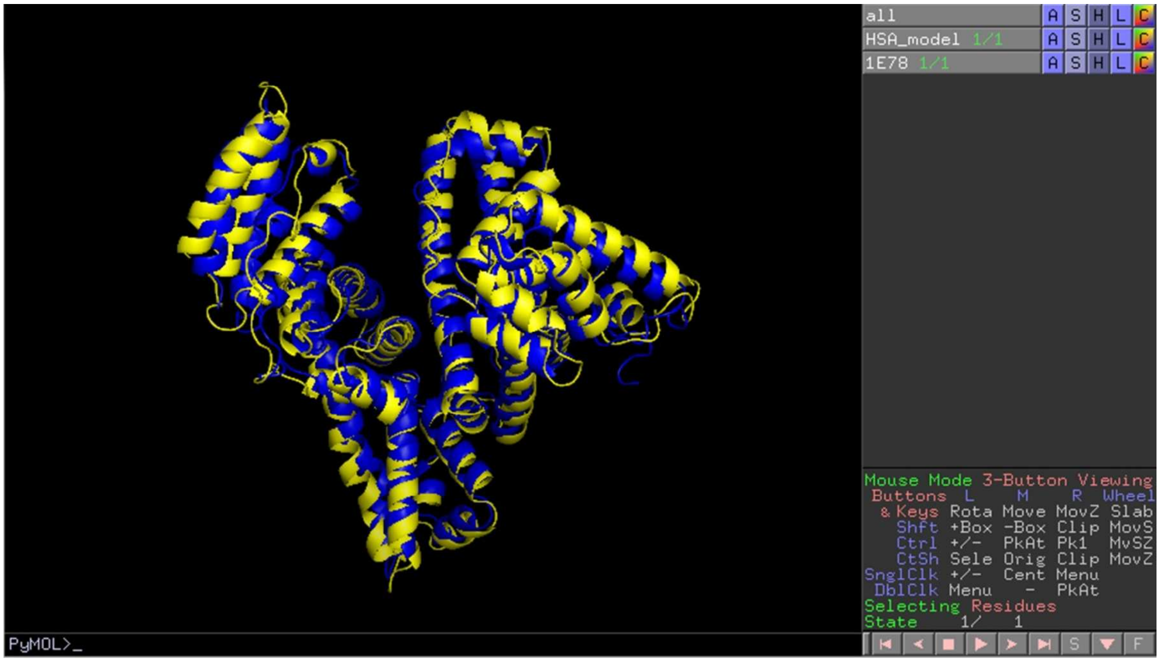


```
PyMOL>_
```

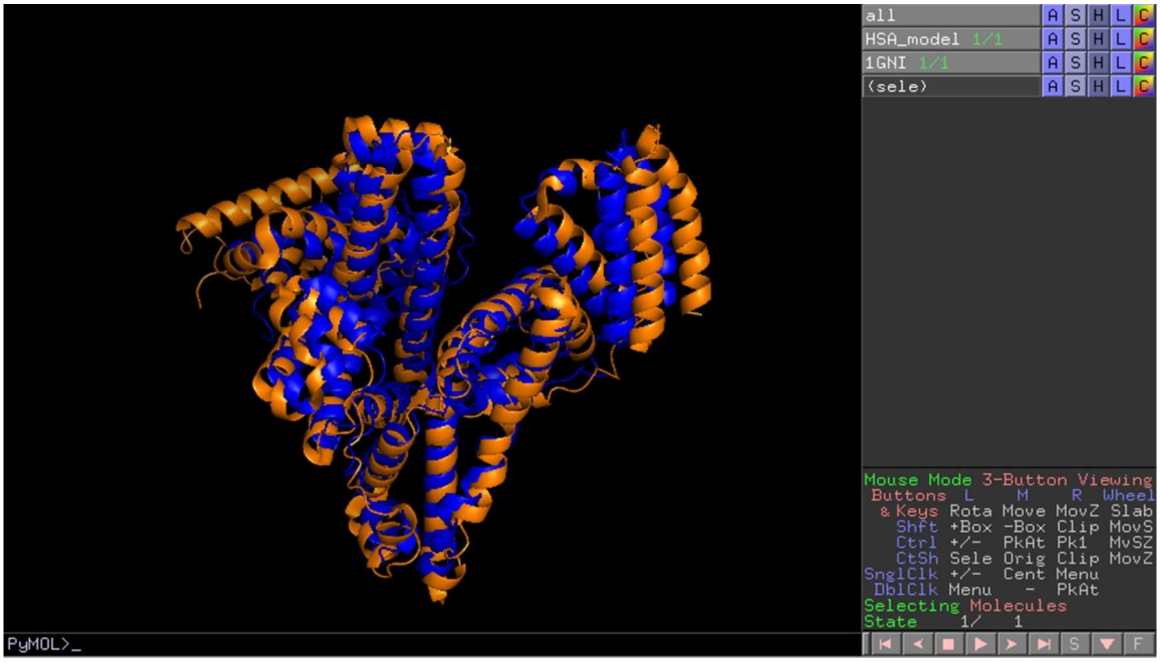
To calculate RMSD$_{\text{all\_atom}}$ value, in terminal window (pres Esc to open/close it) type command:

```
align HSA_model, 1E78
```

and you will get RMSD$_{\text{all\_atom}}$ = 1.770 Å.



To calculate $\text{RMSD}_{\text{all\_atom}}$ value between HSA_model and some other PDB deposited structure open two files (HSA_model and for example 1GNI, co-crystallized with cis-9-octadecenoic acid) and color them differently (HSA_model to blue and 1GNI to orange). After aligning you will get $\text{RMSD}_{\text{all\_atom}}$ = 3.857 Å.



As the previous examples showed, crystal structures can differ more or less in the presence and absence of co-crystallized ligands. Protein tertiary structures and conformational flexibility are highly affected; as the ligand binds the thermal stability of serum albumin usually increases. Therefore, one should always have in mind whether modeling of the active form of enzyme (without inhibitor) or the inhibited form is performed. Also, some enzymes work as holoenzymes: they are active only when an apoenzyme (the protein component of an enzyme) and a coenzyme (a non-protein organic substance) are present. In these cases, one should have a clear idea whether holoenzyme or apoenzyme is being modeled.

PART VIII: FOR INSTRUCTORS & QUESTIONS FOR STUDENTS

HAZARDS

There are no hazards involved with this experiment.

NOTES FOR INSTRUCTOR

The purpose of this exercise is to further students' skills in bioinformatics tools, as well as strengthen students' understanding of protein tertiary structures. Students may perform this laboratory experiment on any computer with internet access and installed educational-use-only PyMOL version (freely available at http://pymol.org/educational/).

Depending on instructor's experience and curriculum of the course involving this laboratory experiment, some other software tools can be used. Further suggestions are given in the table below.

Software tool	**Availability**	**Where to get**
UCSF Chimera	Free for academic use	http://www.cgl.ucsf.edu/chimera/
VMD	Free for academic use	http://www.ks.uiuc.edu/Research/vmd/
Maestro	Free for academic use	http://www.schrodinger.com/
DS Visualizer	Free for academic use	http://accelrys.com/

You may also try to use other Blast servers such as http://mrs.cmbi.ru.nl/mrs-web/blast.do and discuss obtained results with students.

Other homology modeling experiments already exist online (several web locations are listed among references in the manuscript). Depending on a desired level of experiment and scope of learning goals, previous knowledge and experience of students attending this experiment and time available for completion, instructor can decide to use some other protein example than HSA, or even to give a different assignment to every student. Some of the possibilities are given in the table below:

Name of the protein	**UniProt identifier**	**Reference**
tumor necrosis factor ligand superfamily member 6	P41047	Swiss-PdbViewer - Tutorial: Homology Modelling http://spdbv.vital-it.ch/modeling_tut.html
bacterial methylpurine-DNA glycosylase	Q2PAD8	Homology Modeling. http://edu.isb-sib.ch/file.php/57/HM.htm
human Cyclin A1	P78396	Homology Modeling. http://edu.isb-sib.ch/file.php/57/HM.htm
putative protein kinase C delta from Drosophila	P83099	Homology Modeling. http://edu.isb-sib.ch/file.php/57/HM.htm
protein LAP2	Q96RT1	Homology Modeling. http://www.cs.huji.ac.il/~fora/81855/exercises/ex6.pdf

In case of advanced course, we suggest the modeling of one of the proteins from G-protein coupled receptors family, like rhodopsin or beta-adrenergic receptor. These examples would require additional discussion covering specificity and problems of transmembrane receptor modeling and loop predictions.

This laboratory experiment is designed for introductory undergraduate level course of molecular modeling. As this course is mainly designed for students pursuing degree in experimental organic chemistry, highly theoretical background was omitted. The idea of both the course and this experiment was not to educate students to be molecular modelers; rather it was to provide the knowledge of the basic techniques in molecular modeling. For this, simple HSA model was used. Based on the previous knowledge and theoretical background, a short discussion about role of HSA in organism can be involved. In addition, further discussion on properties of conserved and variable regions in homology modeling can be included in the lab experiment, with the particular role of these regions in albumins, both in modeling and in protein activity. Also, it is necessary to underline the structural differences between proteins with and without bound ligands, the role of water in protein shape and function and, if needed, to elaborate on each of the mentioned topics. In this particular course, most of those themes were already covered in introductory theoretical lectures and at the introductory Biochemistry course students attended previously. In the case of the larger groups, with instructor to student ratio higher than 1:15 (our estimation), some additional time may be needed as well. In those cases, one of the possibilities is that final parts of the experiment and some of the questions are given in a form of homework.

All calculations are performed at the basic level, with default settings for majority of used programs. This level is adequate for teaching the major concepts of introductory bioinformatics and homology modeling. However, the instructor and students should discuss about the alignment adjustment during the aligning section, techniques of loop modeling during the modeling phase and structure relaxation and molecular dynamics during the validation phase. Instructor may ask students to experiment with different settings and to compare the obtained results at the end. Also, since all four HSA homologous proteins (ESA, LSA, RSA, and BSA) are very similar to the HSA, instructor may ask students to divide templates among themselves, to prepare homology models based on different templates, and to compare results between themselves at the end.

Students should be reminded that there are other, more advantageous homology modeling software that can give slightly different results. Also, it is a common procedure to relax both crystal structures and prepared homology models using molecular dynamics (MD) simulations. It should be wise to devote one lab class to perform MD of these structures and to analyze differences between optimized and non-optimized structures. If MD lab class cannot be organized, instructor is advised to perform MD simulations of the HSA PDB entry and HSA homology model and to provide students with these two PDB files for subsequent comparison.

Some questions included below may be useful for in-class discussions or for lab reports.

QUESTIONS FOR STUDENTS:

1. In PART I decision was made to use only mature protein (585 residues) and not the complete sequence (609 residues). Provide reasoning for this decision.

> **A:** One out of three proteins is meant to work outside of the cytosol. In order to be transported through the membrane, proteins are synthesized with a short signal peptide. However, for protein to be active, both signal and propeptide sequences have to be cleaved. Since mature protein is the active form – it is the most suitable form for experiments and modeling.

2. In PART II only structures with similarity to HSA of ~75% or more were used. Can structures with smaller similarity percentage be also used and how will it affect the final results?

> **A:** Sequence similarity of more than 50% is generally required, although similarity of more than 30% can be used under certain circumstances. However, in this lab only highly similar structures were used as they provide the best models. Other structures had similarity less than 50% (and very high E-values) so they would contribute only to poor quality models.

3. In PART III the sequence alignment is colored. Based on your knowledge of the standard amino acid structures try to make an educated guess how colors (red, blue, magenta and green) are connected to the following properties: (a) alkaline; (b) acidic; (c) hydroxyl, sulfhydryl, and amine group and (d) small and hydrophobic.

> **A:**
> (a) alkaline = magenta
> (b) acidic = blue
> (c) hydroxyl, sulfhydryl, and amine group = green
> (d) small and hydrophobic = red

4. In PART III below the alignment consensus symbols appear. Based on your knowledge of the standard amino acid structures and their properties try to make an educated guess of the asterisk (*), colon (:) and period (.) meaning.

A:
asterisk = fully conserved residue
colon = residues with highly similar properties
period = residues with slightly similar properties

5. In PART IV crystal structure with PDB ID 4F5U was selected. Among four ESA structures 4F5U has the highest resolution (2.04 Å). What does the resolution tell us about the quality of the crystal structure? What are the other parameters affecting the quality of the structure?

A: Resolution represents the quality of the data obtained from the crystal. It is the measure of details present in the diffraction pattern and electron density map. In excellent crystal structures (high-resolution, about 1 Å) every atom can be easily seen in the electron density map, while in the lower resolution maps (more than 3 Å) it starts to be hard to spot anything more than contours of the protein. Beside resolution, there are other aspects affecting the quality of the crystal structure: R-value, R-free, missing coordinates and missing residues, and others.

6. In PART V when a homology model was made, SWISS MODEL reported Global Model Quality Estimation (GMQE) and QMEAN values. Search the literature to find out which information are these scores providing.

A: GMQE value estimates a quality of target-template alignment and hence expected accuracy of a model; the scale is from 0 to 1, where higher number correlate to higher reliability of a model. QMEAN is a scoring function that estimates the model quality based on four structural descriptors: torsion angles, all-atom interactions, C-beta interactions and solvation.

7. In PART VI a Ramachandran plot was created. Provide your understanding of the homology model quality based on the plotted Φ and Ψ angles.

A: Based on the Ramachandran plot analysis, a high quality homology model was created. Around 97.6% of residues are found in favored region while 2.1% of residues are in allowed region. Only two residues are in the outlier region.

8. In PART VI VADAR analysis of the homology model was performed. Comment on the agreement of observed and expected values of hydrogen bonds, dihedral angles, accessible surface area and volume. Why can certain disagreements with the expected values be tolerated?

> **A:** Although observed number of H-bonds is slightly higher than expected, mean H-bond distance and energy are in good agreement with the expected values. Mean dihedral angles are in good agreement with the expected values. Total accessible surface area is somewhat higher than expected, mainly due to charged residues. Total volume is slightly lower than expected, possibly indicating tighter packing due to higher number of H-bonds. Expected values are idealized, and they are obtained as mean values of different proteins. Since each protein is unique, certain deviations from expected values are allowed. Furthermore, the homology model can be relaxed in an MD simulation to produce much more realistic structure.

9. In PART VII RMSD values between HSA model and two different PDB entries were calculated. Comment on the difference between the two RMSD values, and the effect of the present ligand on the 3D structure of the protein.

> **A:** In order to bind a ligand, protein usually has to undergo some structural changes. When compared to the ligand-free HSA, homology model showed relatively small RMSD of 1.77 Å. However, when compared to the HSA bound to cis-9-octadecenoic acid, homology model showed an increase in RMSD to 3.86 Å. Therefore, we can conclude that structural differences between ligand-free and ligand-bound HSA exist.

10. Comment on the similarities and differences in HSA model before and after molecular dynamics optimization. Which structure is more realistic according to the Ramachandran plot and VADAR analysis? Why?

> **A:** MD simulation was not run as a part of this lab experiments. However, structure should get more realistic after MD simulation. Therefore, both Ramachandran plot and VADAR analysis should show this. During an MD simulation protein is allowed to relax, over the time, in its natural environment. Furthermore, proteins are not static structures so their properties are much better explained under dynamic conditions.