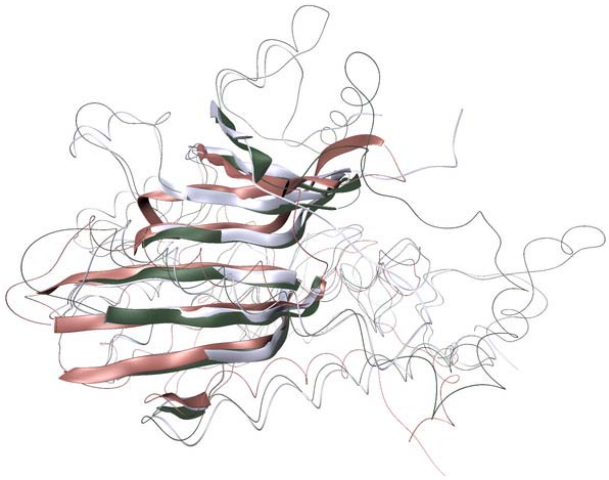
University of Illinois at Urbana-Champaign
Department of Chemistry, Luthey-Schulten Group
Beckman Institute for Advanced Science and Technology
Theoretical and Computational Biophysics Group
Summer School 2003

Bioinformatics, Sequence and Structural Alignment



Felix Autenrieth

Barry Israelewitz

Zaida Luthey-Schulten

Anurag Sethi

Taras Pogorelov

June 2003

A web version, in color, is available at
http://www.ks.uiuc.edu/Training/SumSchool03/Tutorials/bioinformatics

Contents

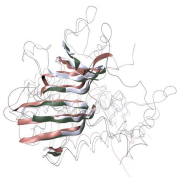
Introduction

The recent developments of projects such as the sequencing of the genome from several organisms, and high-throughput X-ray structure analysis, have brought to the scientific community a large amount of data about the sequences and structures of several thousand proteins. This information can effectively be used for medical and biological research only if one can extract functional insight from the sequence and structural data. To achieve this we need to understand how the proteins perform their functions. Two main computational techniques exist to reach this a goal: a *bioinformatics* approach, and atomistic *molecular dynamics* simulations. Bioinformatics uses the statistical analysis of protein sequences and structures to understand their function and predict structures when only sequence information is available. Molecular modeling and molecular dynamics simulations use the principles from physics and physical chemistry to study the function and folding of proteins.

Bioinformatics methods are among the most powerful technologies available in life sciences today. They are used in fundamental research on theories of evolution and in more practical considerations of protein design. Algorithms and approaches used in these studies range from sequence and structure alignments, secondary structure prediction, functional classification of proteins, threading and modeling of distantly-related homologous proteins to modeling the progress of protein expression through a cell's life cycle.

In this tutorial you will use classical sequence alignment methods; with the Smith-Waterman [1] and Needleman-Wunsch algorithms. You will start out only with sequence and biological information of class II aminoacyl-tRNA synthetases, key players in the translational mechanism of cell. Then you will classify protein domains and align the determined protein domains structurally. If structural alignments are considered to be the true alignments, you will see that simple pair sequence alignment of two proteins with low sequence identity has serious limitations. Finally you will determine the phylogenetic relationship of class II tRNA synthetases with a dendrogram-creation algorithm. You will carry out the exercises with the programs VMD, MOE, MATLAB and a Needleman-Wunsch alignment program provided by A. Sethi. Many of the tools of the field can be freely accessed by any person with a web browser; a listing of our favorite bioinformatics tools and resources is provided.

The entire tutorial takes about 2 hours to complete.



Protein sequences vs. nucleotide sequences. A protein is a sequence of amino acids linked with peptide bonds to form a polypeptide chain. In this tutorial, the word *sequence* (unless we note otherwise) refers to the amino acid residue sequence of a protein; by convention these sequences are listed from the N-terminal to the C-terminal of the chain. Sequences can be written with full names, as in "Lysine, Arginine, Cytsteine, ...", with 3-letter codes, "Lys, Arg, Cys, ...", or with 1-letter codes, "K, R, C, ..." (as listed in Table 1). Proteins range in size from a few dozen to several thousand residues. The nucleotide sequences of DNA encodes protein sequence. Sections of genes in chromosomal DNA are copied to mRNA, which provides the guide for ribosome to assemble a protein. A nucleotide sequence may be written as "Cytosine, Adenine, Adenine, Guanine, ...", or "C, A, A, G, ...".

This tutorial assumes that VMD, Moe, Matlab and other software has been correctly installed on the user's computer. Please ask a lab attendant for help if you have any trouble locating software or data files during the tutorial.

To set up the exercises...

You will make a copy of the files needed for these exercises in your home directory. Open a terminal window, and, if you don't already have one, make a `~/tbss.work` directory by typing at the Unix prompt:
>> `mkdir ~/tbss.work`
Make sure that you have a `~/tbss.work` directory:
>> `ls ~/tbss.work`

Copy the needed directory, but instead of typing *TOP_DIR*, type the location of the Summer School directory tree:
>> `cp -rp` *TOP_DIR*`/sumschool03/tutorials/07-bioinformatics/files/Bioinformatics/ ~/tbss.work/`
For instance, if the materials are located at `/mnt/cdrom`, you will type:
`cp -rp /mnt/cdrom/sumschool03/tutorials/07-bioinformatics/files/Bioinformatics/ ~/tbss.work/`

Check that you have the files in this directory by listing the contents:
>> `cd ~/tbss.work/Bioinformatics`
>> `ls -lR`

In this tutorial, when we refer to `~/tbss.work/Bioinformatics/` and its subdirectories, we are referring to the copy which you have just made in your own home directory.

1 Biology of class II aminoacyl-tRNA Synthetases

Translation in biological cells is the process of protein synthesis directed by a nucleic acid message, mRNA. In the ribosome, each set of three successive nucleotides in the mRNA is matched to a specific amino acid according to the code shown in Table 2. The translation machinery dedicated to interpreting this nucleic acid code operates in a two part process. Amino acids are covalently linked, or "charged", to their cognate transfer RNAs (tRNAs) via an aminoacylation reaction catalyzed by a diverse group of multi-domain proteins, the aminoacyl-tRNA synthetases. Charged tRNAs are then shuttled to the ribosome where the tRNA anti-codon is matched to the mRNA codon, and the tRNA is deacylated with the amino acid being added as the next residue of a nascent protein chain [2]. The RNA world hypothesis states that the modern biological world – which relies on DNA and RNA to store genetic information and on proteins to perform catalytic tasks – was pre-dated by and evolved from a form of life that was mostly RNA based, with RNA molecules serving not only to store information, but also to perform required catalytic functions. It is likely that among the first proteins to take over catalytic duties from ribozymes were the aminoacyl-tRNA synthetases (AARSs). These ancient proteins are found in all extant organisms, and their inception clearly pre-dates the root of the universal phylogenetic tree [3, 4]. In this tutorial you will use several alignment methods to study and compare various AARSs.

Amino Acid	Single Letter	Three Letter
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic acid	D	Asp
Asparagine or aspartic acid	B	Asx
Cysteine	C	Cys
Glutamine	Q	Gln
Glutamic acid	E	Glu
Glutamine or glutamic acid	Z	Glx
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val

Table 1: Amino acids names and letter codes

	U		C		A		G		
	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
U	UUC	Phe	UCC	Ser	UAC	Tory	GU	Cys	C
	UUA	Leu	UCA	Ser	UAA	*stop*	UGA	*stop*	A
	UUG	Leu	UCG	Ser	UAG	*stop*	UGG	Trp	G
	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG*	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
G	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG*	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table 2: The genetic code. Some species have slightly different codes. *This codon also specifies the initiator $tRNA^{fMet}$.

2 AARSs in *Methanococcus jannaschii*

In this section, you will find all the tRNA synthetases in an organism. The M. janaschii *genome [5], with NCBI accession number* NC_000909*, has been completely sequenced, so we can perform comprehensive searches through all of its genes.*

- Open the Mozilla web browser: type `mozilla` at a command line or click on the appropriate icon.
- Access the NCBI database (`http://www.ncbi.nlm.nih.gov`).
- Type the *M. jannaschii* accession number NC_000909 into the Search box, with Genome selected as the search type, click on the single result, and you will reach the site with the complete annotated sequence of *M. jannaschii*.
- The best way to find Class II tRNA synthetases in this genome is via the listing of proteins organized by COG (Cluster of Orthologous Groups) functional categories. Click on the COG functional categories link located about halfway down the page as part of the text "Gene Classifaction based on **COG functional categories**" (note: this is *not* the COGs link, located near the top of the page, just beneath "BLAST protein homologs"). You will reach a catalog of *M. jannaschii* proteins organized by functional annotation.
- Choose the link which it most likely to provide you with information about class II tRNA synthetases. (Hint: tRNA synthetase function is part of *translation*.) Click on this link and scroll to the search fields at the bottom of the resulting page.
- Type the string `tRNA synthetase` in the text search box and you will receive a summary of all tRNA synthetases in *M. jannaschii*. Consult the genetic code in Table 2 for help in answering the following questions.

Questions. How many tRNA synthetases are produced by *M. janaschii*? Among these, are there any that bind the same amino acid? Why might this be the case? How many tRNA synthetases are minimally required to synthesize all proteins in one organism? What is "codon usage" and how can information about codon usage be applied to distinguish genomes from two organisms?

3 Domain structure of class II tRNA synthetases

In this section you will study the similarity between functional domains in tRNA synthetases from two domains of life.

In the following steps, we will perform a search for the tRNA synthetases from *E. coli* and *M. jannaschii* in a database that can perform searches based on similar domain structure, the NCBI Entrez Structure/MMDB/3D Domain tools. Sample output is shown in Figure 1.

- At the home page of the NCBI database (`http://www.ncbi.nlm.nih.gov`), set the search-type pop-up menu to Structure and search for `tRNA synthetase`.
- In the search result lists, click on the links for the individual tRNA synthetases (e.g. 1NJ5) to access information about the domain structure of individual class II tRNA synthetases. Click on individual domains in the 3D Domain to see a display of structural neighbors: domains with similar structure to your target domain are shown aligned to your target domain. (Warning: if you receive the error "Vast neighbor data for this domain are not yet available.", just use the Back button to return to the search results and try again with another structure.)
- The online web version of this tutorial includes a color figure of the domains of tRNA synthetase, color-coded by domain as Figure 1. The NCBI domain results are colored by the same code as in this figure: magenta= Catalytic, blue= Insert I, orange= Insert II, green = Anticodon.
- Look among the entires in your search results to find tRNA synthetases from two or three different species. Open separate browser windows to show the domain organization of each and compare their 1-D organization.
- Look among the entires in your search results to find two or three tRNA synthetases from the same species. Open separate browser windows to show the domain organization of each and compare their 1-D organization.

Question. What is the biological function of protein domains in one of your chosen tRNA synthetases? Which is more similar: the domain structures of different tRNA synthetases within one organism (paralogs), or the domain structures of tRNA synthetases from different organisms (orthologs)?

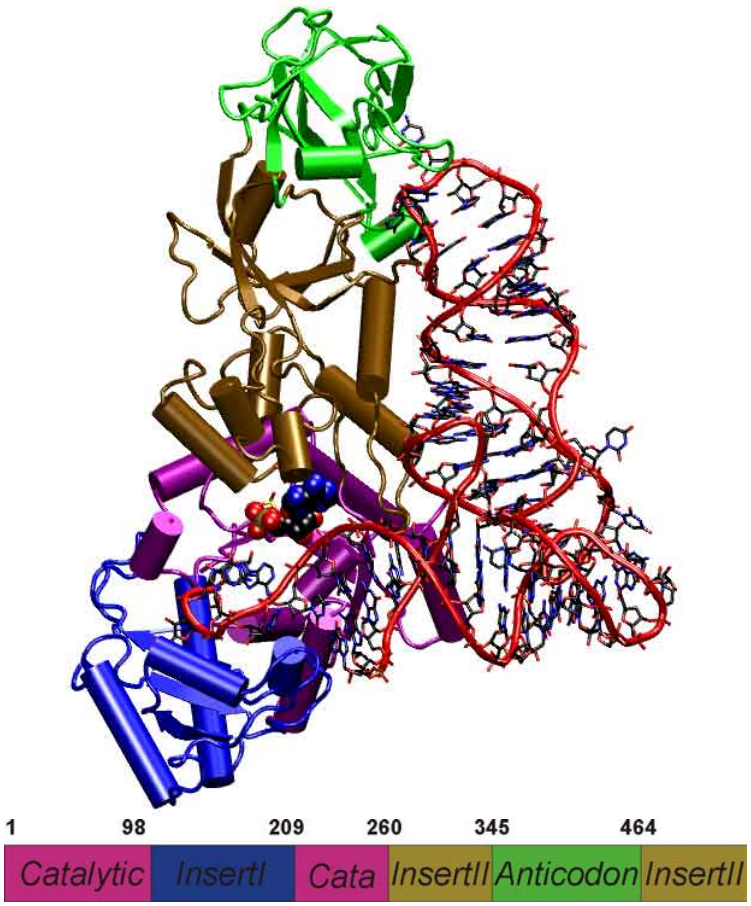


Figure 1: Domains of Glutamine aminoacyl-tRNA synthetase. The bound tRNA is shown with red backbone on the right side of the figure.

4 SCOP fold classification

As with all multi-domain proteins, the properties of Class II tRNA synthetases depend on their component domains. The schematic domain structure of aspartyl class II tRNA synthetase from Eubacteria is depicted in Figure 1.

Only one domain is common to all class II tRNA synthetases studied so far: the catalytic domain, which carries out the amino acid loading reaction. In later sections of this tutorial, you will align the catalytic domains of class II tRNA synthetases using Needleman-Wunsch sequence alignment, and with structure-based alignment methods. The fourth column of Table 6 provide you with the ASTRAL database accession codes [6]. The ASTRAL database (`http://astral.stanford.edu`) is a compendium of protein domain structures derived from the PDB database [7].

In the following exercise you will classify one of your chosen catalytic domains in folds, superfamilies and families with help of the database SCOP (Structural Classification Of Proteins) [8].

- Point your web browser to the SCOP server (`http://scop.berkeley.edu/index.html`).
- You are going to browse a hierarchy of protein structural classifications. Go to the top of the hierarchy: Click on the link top of the hierarchy. This places you at the root of all the protein classes deposited in SCOP.
- Click on one of the Classes links (e.g. Alpha and Beta proteins); you should reach the next hierarchy level of folds. Note how lineage records your path through the hierarchy.
- Some of the entries have pre-prepared 3D structure renderings of an example of the protein class. Click on the purple and white buttons to the right of a few entries to see an example view of the chosen protein class.
- We will now search for entries relevant to a particular domain of interest. Now enter one of the ASTRAL database catalytic domain codes from Table 6 (e.g. `d1asza2`) in the Search field you find at the bottom of the page.
- You should reach a window showing the SCOP lineage and a summary of all relevant PDB entry domains. For subsequent exercises we have provided you with catalytic domain coordinates of Class II tRNA synthetases obtained from the ASTRAL database. The ASTRAL coordinate files are produced by extracting domains from PDB coordinate files.

Questions. What is the lineage of your chosen tRNA synthetase domain? What are some other members in the SCOP family and superfamily of the catalytic domain of Class II tRNA synthetase? What is the most abundant fold in SCOP's "Alpha and Beta proteins" fold class ?

5 Sequence Alignment Algorithms

In this section you will optimally align two short protein sequences using pen and paper, then search for homologous proteins by using a computer program to align several, much longer, sequences.

Dynamic programming algorithms are recursive algorithms modified to store intermediate results, which improves efficiency for certain problems. The Smith-Waterman (Needleman-Wunsch) algorithm uses a dynamic programming algorithm to find the optimal local (global) alignment of two sequences — a and b. The alignment algorithm is based on finding the elements of a matrix H where the element $H_{i,j}$ is the optimal score for aligning the sequence (a_1,a_2,...,a_i) with (b_1,b_2,.....,b_j). Two similar amino acids (e.g. arginine and lysine) receive a high score, two dissimilar amino acids (e.g. arginine and glycine) receive a low score. The higher the score of a path through the matrix, the better the alignment. The matrix H is found by progressively finding the matrix elements, starting at $H_{1,1}$ and proceeding in the directions of increasing i and j. Each element is set according to:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

where $S_{i,j}$ is the similarity score of comparing amino acid a_i to amino acid b_j (obtained here from the BLOSUM40 similarity table) and d is the penalty for a single gap. The matrix is initialized with $H_{0,0} = 0$. When obtaining the local Smith-Waterman alignment, $H_{i,j}$ is modified:

$$H_{i,j} = \max \begin{cases} 0 \\ H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

The gap penalty can be modified, for instance, d can be replaced by $(d \times k)$, where d is the penalty for a single gap and k is the number of consecutive gaps.

Once the optimal alignment score is found, the "traceback" through H along the optimal path is found, which corresponds to the the optimal sequence alignment for the score. In the next set of exercises you will manually implement the Needleman-Wunsch alignment for a pair of short sequences, then perform global sequence alignments with a computer program developed by Anurag Sethi, which is based on the Needleman-Wunsch algorithm with an affine gap penalty, $d + e(k-1)$, where e is the extension gap penalty. The output file will be in the GCG format, one of the two standard formats in bioinformatics for storing sequence information (the other standard format is FASTA).

5.1 Manually perform a Needleman-Wunsch alignment

In the first exercise you will test the Smith-Waterman algorithm on a short sequence parts of hemoglobin (PDB code `1AOW`) and myoglobin 1 (PDB code

`1AZI`).

- Here you will align the sequence `HGSAQVKGHG` to the sequence `KTEAEMKASEDLKKHGT`.
- The two sequences are arranged in a matrix in Table 3. The sequences start at the upper right corner, the initial gap penalties are listed at each offset starting position. With each move from the start position, the initial penalty increase by our single gap penalty of 8.

		H	G	S	A	Q	V	K	G	H	G
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
K	-8										
T	-16										
E	-24										
A	-32										
E	-40										
M	-48										
K	-56										
A	-64										
S	-72										
E	-80										
D	-88										
L	-96										
K	-104										
K	-112										
H	-120										
G	-128										
T	-136										

Table 3: The empty matrix with initial gap penalties.

- The first step is to fill in the similarity scores $S_{i,j}$ from looking up the matches in the BLOSUM40 table, shown here labeled with 1-letter amino acid codes:

```
A  5
R -2  9
N -1  0  8
D -1 -1  2  9
C -2 -3 -2 -2 16
Q  0  2  1 -1 -4  8
E -1 -1 -1  2 -2  2  7
G  1 -3  0 -2 -3 -2 -3  8
H -2  0  1  0 -4  0  0 -2 13
I -1 -3 -2 -4 -4 -3 -4 -4 -3  6
L -2 -2 -3 -3 -2 -2 -2 -4 -2  2  6
K -1  3  0  0 -3  1  1 -2 -1 -3 -2  6
M -1 -1 -2 -3 -3 -1 -2 -2  1  1  3 -1  7
F -3 -2 -3 -4 -2 -4 -3 -3 -2  1  2 -3  0  9
P -2 -3 -2 -2 -5 -2  0 -1 -2 -2 -4 -1 -2 -4 11
S  1 -1  1  0 -1  1  0  0 -1 -2 -3  0 -2 -2 -1  5
T  0 -2  0 -1 -1 -1 -1 -2 -2 -1 -1  0 -1 -1  0  2  6
W -3 -2 -4 -5 -6 -1 -2 -2 -5 -3 -1 -2 -2  1 -4 -5 -4 19
Y -2 -1 -2 -3 -4 -1 -2 -3  2  0  0 -1  1  4 -3 -2 -1  3  9
V  0 -2 -3 -3 -2 -3 -3 -4 -4  4  2 -2  1  0 -3 -1  1 -3 -1  5
B -1 -1  4  6 -2  0  1 -1  0 -3 -3  0 -3 -3 -2  0  0 -4 -3 -3  5
Z -1  0  0  1 -3  4  5 -2  0 -4 -2  1 -2 -4 -1  0 -1 -2 -2 -3  2  5
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1  0 -1 -2  0  0 -2 -1 -1 -1 -1 -1
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X
```

- We fill in the BLOSUM40 similarity scores for you in Table 4.
- To turn this S matrix intro the dynamic programming H matrix requires calculation of the contents of all 170 boxes. We've calculated the first 4 here, and encourage you to calculate the contents of at least 4 more. The practice will come in handy in the next steps. As described above, a matrix square cannot be filled with its dynamic programming value until the squares above, to the left, and to the above-left diagonal are computed. The value of a square is

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

  , using the convention that H values appear in the top part of a square in large print, and S values appear in the bottom part of a square in small print. Our gap penalty d is 8.
- Example: In the upper left square in Table 4, square (1,1), the similarity score $S_{1,1}$ is -1, the number in small type at the bottom of the box. The value to assign as $H_{1,1}$ will be the greatest ("max") of these three values: $(H_{0,0}+S_{1,1}), (H_{0,1}-d), (H_{1,0}-d))$. That is, the greatest of: $(0+-1), (-8-8), (-8-8)$ which just means the greatest of: -1, -16, and -16. This is -1, so we write -1 as the value of $H_{1,1}$ (the larger number in the top part of the box). The same reasoning in square (2,1) leads us to set $H_{2,1}$ as -9, and so on.

  Note: we consider $H_{0,0}$ to be the "predecessor" of $H_{1,1}$, since it helped decided $H_{1,1}$'s value. Later, predecessors will qualify to be on the traceback path.
- Again, just fill in 4 or 5 boxes in Table 4 until you get a feel for gap penalties and similarity scores S vs. alignment scores H. In the next step, we provide the matrix with all values filled in as Table 5. Check that your 4 or 5 calculations match.
- Now we move to Table 5, with all 170 $H_{i,j}$ values are shown, to do the "alignment traceback". To find the alignment requires one to trace the path through from the end of the sequence (the lower right box) to the start of the sequence (the upper left box). This job looks complicated, but should only take about 5 –7 minutes.
- We are tracing a path in Table 5, from the lower right box to the upper left box. You can only move to a square if it could have been a "predecessor" of your current square – that is, when the matrix was being filled with $H_{i,j}$ values, the move from the predecessor square to your current square would have followed the mathematical rules we used to find $H_{i,j}$ above. Circle each square you move to along your path.

- Example: we start at the lower right square (10,17), where $H_{10,17}$ is -21 and $S_{10,17}$ is -2. We need to test for 3 possible directions of movement: diagonal (up + left), up, and left. The condition for diagonal movement given above is: $H_{i,j} = H_{i-1,j-1} + S_{i,j}$, so for the diagonal box (9,16) to have contributed to (10,17), $H_{9,16} + S_{10,17}$ would have to equal the H value of our box, -21. Since (-29 + -2) does not equal -21, the diagonal box is not a "predecessor", so we can't move in that direction. We try the rule for the box to the left: $H_{i,j} = H_{i-1,j} - d$ Since -37 - 8 does not equal -21, we also can't move left. Our last chance is moving up. We test $H_{i,j} = H_{i,j-1} - d$. Since -21 = (-13 - 8) we can move up! Draw an arrow from the lower right box, $(H_{10,17} = -21, S_{10,17} = -2)$ to the box just above it, $(H_{10,16} = -13, S_{10,16} = 8)$.

- Continue moving squares, drawing arrows, and circling each new square you land on, until you have reached the upper right corner of the matrix If the path branches, follow both branches.

- Write down the alignment(s) that corresponds to your path(s) by writing the the letter codes on the margins of each position along your circled path. Aligned pairs are at the boxes at which the path exits via the upper-left corner. When there are horizontal or vertical movements movements along your path, there will be a gap (write as a dash, "-") in your sequence.

- Now to check your results against a computer program. We have prepared a pairwise Needleman-Wunsch alignment program, `pair`, which you will apply to the same sequences which you have just manually aligned.

- Change your directory by typing at the Unix prompt:
  `cd ~/tbss.work/Bioinformatics/pairData`
  then start the pair alignment executable by typing:
  `pair targlist`
  . All alignments will be carried out using the BLOSUM40 matrix, with a gap penalty of 8. The paths to the input files and the BLOSUM40 matrix used are defined in the file `targlist`; the BLOSUM40 matrix is the first 25 lines of the file `blosum40`. (Other substitution matrices can be found at the NCBI/Blast website.)

  Note: In some installations, the `pair` *executable is in* `~/tbss.work/Bioinformatics/pairData` *and here you must type* `./pair targlist` *to run it.*
  If you cannot access the `pair` *executable at all, you can see the output from this step in* `~/tbss.work/Bioinformatics/pairData/example_output/`

- After executing the program you will generate three output files namely `align`, `scorematrix` and `stats`. View the alignment in GCG format by typing `less align`. The file `scorematrix` is the 17x10 H matrix. If there are multiple paths along the traceback matrix, the program `pair`

will choose only one path, by following this precedence rule for existing potential traceback directions, listed in decreasing precedence: diagonal (left and up), up, left. In the file `stats` you will find the optimal alignment score and the percent identity of the alignment.

Questions. Compare your manual alignment to the the output of the pair program. Do the alignments match?

5.2 Finding homologous pairs of ClassII tRNA synthetases

Homologous proteins are proteins derived from a common ancestral gene. In this exercise with the Needleman-Wunsch algorithm you will study the sequence identity of several class II tRNA synthetases, which are either from Eucarya, Eubacteria or Archaea or differ in the kind of aminoacylation reaction which they catalyze. Table 6 summarizes the reaction type, the organism and the PDB accession code and chain name of the employed Class II tRNA synthetase domains.

- We have have prepared a computer program `multiple` which will align multiple pairs of proteins.
- Change your directory by typing at the Unix prompt:
  `cd ~/tbss.work/Bioinformatics/multipleData`
  then start the alignment executable by typing:
  `multiple targlist`
  .
  Note: In some installations, the `multiple` *executable is in* `~/tbss.work/Bioinformatics/multipleData` *and here you must type* `./multiple targlist` *to run it.*
  If you cannot access the `multiple` *executable at all, you can see the output from this step in* `~/tbss.work/Bioinformatics/multipleData/example_output/`
- In the `align` and `stats` files you will find all combinatorial possible pairs of the provided sequences. On a piece of paper, write the names of the the proteins, grouped by ther domain of life, as listed in Table 6. Compare sequence identities of aligned proteins from the same domain of a life, and of aligned proteins from different domains of life, to help answer the questions below.

Questions. What criteria do you use in order to determine if two proteins are homologous? Can you find a pattern when you evaluate percent identities between the pairs of class II tRNA synthetases? Which is the most evolutionarily related pair, and which is the most evolutionarily divergent pair according to the sequence identity?

		H	**G**	**S**	**A**	**Q**	**V**	**K**	**G**	**H**	**G**
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
K	-8	**−1** −1	**−9** −2	0	−1	1	−2	6	−2	−1	−2
T	-16	**−9** −2	**−3** −2	2	0	−1	1	0	−2	−2	−2
E	-24	0	−3	0	−1	2	−3	1	−3	0	−3
A	-32	−2	1	1	5	0	0	−1	1	−2	1
E	-40	0	−3	0	−1	2	−3	1	−3	0	−3
M	-48	1	−2	−2	−1	−1	1	−1	−2	1	−2
K	-56	−1	−2	0	−1	1	−2	6	−2	−1	−2
A	-64	−2	1	1	5	0	0	−1	1	−2	1
S	-72	−1	0	5	1	1	−1	0	0	−1	0
E	-80	0	−3	0	−1	2	−3	1	−3	0	−3
D	-88	0	−2	0	−1	−1	−3	0	−2	0	−2
L	-96	−2	−4	−3	−2	−2	2	−2	−4	−2	−4
K	-104	−1	−2	0	−1	1	−2	6	−2	−1	−2
K	-112	−1	−2	0	−1	1	−2	6	−2	−1	−2
H	-120	13	−2	−1	−2	0	−4	−1	−2	13	−2
G	-128	−2	8	0	1	−2	−4	−2	8	−2	8
T	-136	−2	−2	2	0	−1	1	0	−2	−2	−2

Table 4: Alignment score worksheet. In all alignment boxes, the similarity score $S_{i,j}$ from the BLOSUM40 matrix lookup is supplied (small text, bottom of square). Four alignment scores are provided as examples (large text, top of square), try and calculate at least four more, following the direction provided in the text for calculating $H_{i,j}$.

		H	**G**	**S**	**A**	**Q**	**V**	**K**	**G**	**H**	**G**
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
K	-8	**−1** −1	**−9** −2	**−16** 0	**−24** −1	**−31** 1	**−39** −2	**−42** 6	**−50** −2	**−58** −1	**−66** −2
T	-16	**−9** −2	**−3** −2	**−7** 2	**−15** 0	**−23** −1	**−30** 1	**−38** 0	**−44** −2	**−52** −2	**−60** −2
E	-24	**−16** 0	**−11** −3	**−3** 0	**−8** −1	**−13** 2	**−21** −3	**−29** 1	**−37** −3	**−44** 0	**−52** −3
A	-32	**−24** −2	**−15** 1	**−10** 1	**2** 5	**−6** 0	**−13** 0	**−21** −1	**−28** 1	**−36** −2	**−43** 1
E	-40	**−32** 0	**−23** −3	**−15** 0	**−6** −1	**4** 2	**−4** −3	**−12** 1	**−20** −3	**−28** 0	**−36** −3
M	-48	**−39** 1	**−31** −2	**−23** −2	**−14** −1	**−4** −1	**5** 1	**−3** −1	**−11** −2	**−19** 1	**−27** −2
K	-56	**−47** −1	**−39** −2	**−31** 0	**−22** −1	**−12** 1	**−3** −2	**11** 6	**3** −2	**−5** −1	**−13** −2
A	-64	**−55** −2	**−46** 1	**−38** 1	**−26** 5	**−20** 0	**−11** 0	**3** −1	**12** 1	**4** −2	**−4** 1
S	-72	**−63** −1	**−54** 0	**−41** 5	**−34** 1	**−25** 1	**−19** −1	**−5** 0	**4** 0	**11** −1	**4** 0
E	-80	**−71** 0	**−62** −3	**−49** 0	**−42** −1	**−32** 2	**−27** −3	**−13** 1	**−4** −3	**4** 0	**8** −3
D	-88	**−79** 0	**−70** −2	**−57** 0	**−50** −1	**−40** −1	**−35** −3	**−21** 0	**−12** −2	**−4** 0	**2** −2
L	-96	**−87** −2	**−78** −4	**−65** −3	**−58** −2	**−48** −2	**−38** 2	**−29** −2	**−20** −4	**−12** −2	**−6** −4
K	-104	**−95** −1	**−86** −2	**−73** 0	**−66** −1	**−56** 1	**−46** −2	**−32** 6	**−28** −2	**−20** −1	**−14** −2
K	-112	**−103** −1	**−94** −2	**−81** 0	**−74** −1	**−64** 1	**−54** −2	**−40** 6	**−34** −2	**−28** −1	**−22** −2
H	-120	**−99** 13	**−102** −2	**−89** −1	**−82** −2	**−72** 0	**−62** −4	**−48** −1	**−42** −2	**−21** 13	**−29** −2
G	-128	**−107** −2	**−91** 8	**−97** 0	**−88** 1	**−80** −2	**−70** −4	**−56** −2	**−40** 8	**−29** −2	**−13** 8
T	-136	**−115** −2	**−99** −2	**−89** 2	**−96** 0	**−88** −1	**−78** 1	**−64** 0	**−48** −2	**−37** −2	**−21** −2

Table 5: Traceback worksheet. The completed alignment score matrix H (large text, top of each square) with the BLOSUM40 lookup scores $S_{i,j}$ (small text, bottom of each square). To find the alignment, trace back starting from the lower right (T vs G, score -21) and proceed diagonally (to the left and up), left, or up. Only proceed, however, if the square in that direction could have been a predecessor, according to the conditions described in the text.

Specificity	Organism	PDB code:chain	ASTRAL catalytic domain
Aspartyl	Eubacteria	1EQR:B	d1eqrb3
Aspartyl	Archaea	1B8A:A	d1b8aa2
Aspartyl	Eukarya	1ASZ:A	d1asza2
Glycl	Archaea	1ATI:A	d1atia2
Histidyl	Eubacteria	1ADJ:C	d1adjc2
Lysl	Eubacteria	1BBW:A	d1bbwa2
Aspartyl	Eubacteria	1EFW:A	d1efwa3

Table 6: Domain types, origins, and accession codes

6 Sequence and structural alignments in MOE

In this section we will investigate two catalytic domain pairs of class II tRNA synthetases, by examining their relatedness by means of sequence and structural alignments.

This will show the limitations of simple pairwise sequence alignment methods for sequences with low sequence identity.

More related:
Pair 1 →
`d1eqrb3` *vs.* `d1efwa3` (Aspartyl Eubacteria `1EQR:B` *vs.* Aspartyl Eubacteria `1EFW:A`)

More divergent:
Pair 2 →
`d1eqrb3` *vs.* `d1adjc2` (Aspartyl Eubacteria `1EQR:B` *vs.* Histidyl Archaea `1ADJ:C`)

All our alignments will be carried out with the BLOSUM40 matrix , a gap start penalty of 12 and an extended gap penalty of 2 (you can try other substitution matrices and gap penalties). You will carry out your alignments in the sequence editor of Moe. The sequences will be structurally aligned first by sequence and then using structure-based methods.

6.1 Align Pair 1 by sequence

- Go to the directory `~/tbss.work/Bioinformatics/moeData`
- Start Moe by typing `moe` at the command line.
- Load the two structures of Pair 1 into Moe by clicking on: File:Open in the main Moe menu. Select `d1eqrb3.ent`. Click OK when the "Read PDB file" option box appears". Repeat for `d1ewfa3.ent`.
- After loading the pair, center both of them by clicking on View on the right button bar.

Sequence alignment

We will work from now on only in the Sequence Editor. Select Window:Sequence editor from the main Moe menu. The first alignment will be a sequence alignment with the following setup:

- Click on: Homology:Align in the Sequence Editor menu.
- Following Figure 3, change the substitution matrix to BLOSUM40, the gap start value to 12 and the gap extend value to 2. Uncheck round-robin, iterative refinement, and structural alignment and superpose chains.
- Hit the OK button and Moe will carry out a sequence alignment.

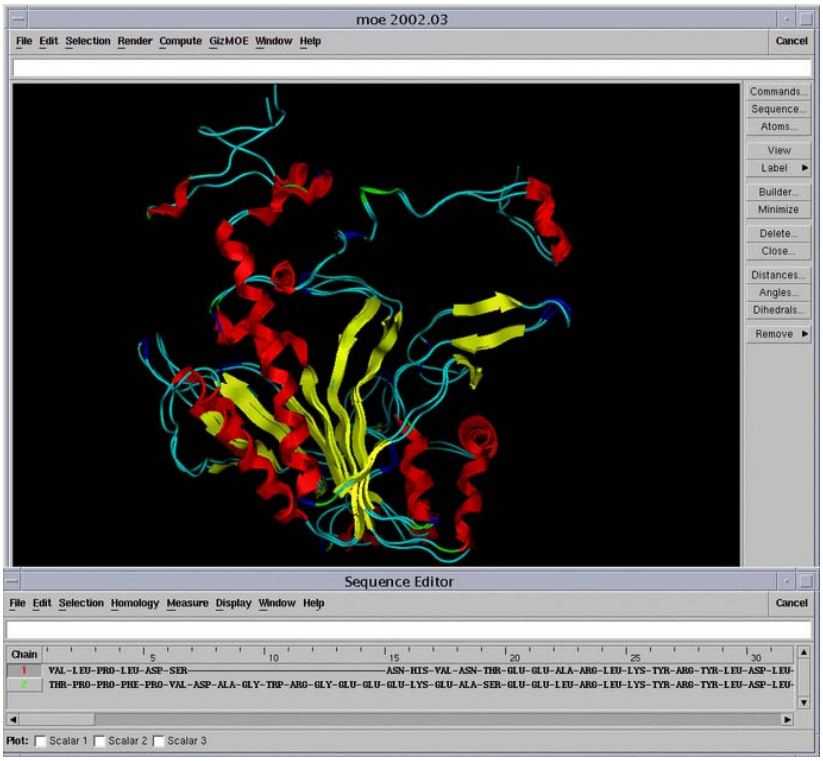


Figure 2: A pair of aligned domains displayed in MOE.

Question. Display alignment using single-letter code by selecting Display:Single Letter Residues. Show names of compound by selecting Display:Compound Name. How do the sequence alignments carried out with the Needleman-Wunsch program `pair` compare to the sequence alignments in MOE?

Superposition

In the next steps you will superpose all CA atoms of both structures according to your sequence alignment

- Click on Homology:Superpose in the Sequence Editor menu.
- Press Superpose.
- Make a note of the RMSD value.

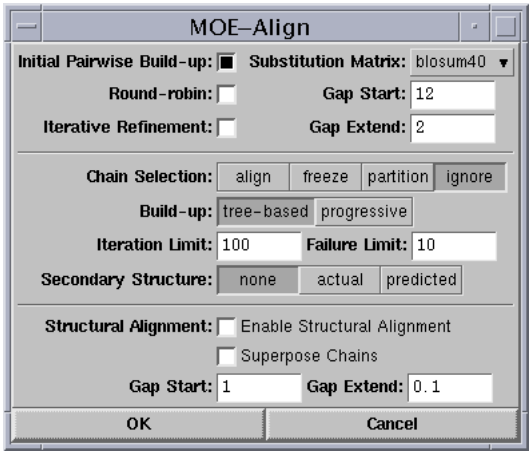


Figure 3: Sequence alignment in MOE.

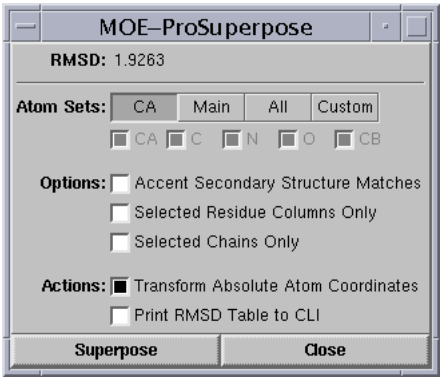


Figure 4: Superpostion options in MOE.

- For better visualization of the superposition render the structures in Cartoon representation and hide all other atom representations. Select Render:Backbone:Cartoon and Render:Hide All from the main menu.
- For rendering and selection commands refer to the "Determining Force Fields" tutorial.

6.2 Align Pair 1 by structure

- Select chain 1 and chain 2 in the sequence window by shift-clicking the square buttons at the left of each sequence. Reset the sequence alignment by clicking on: Homology:Reset Alignment

- Carry out the same steps as for the sequence alignment except this time be sure to check the Structural Alignment boxes as depicted in Figure 5.
- Superimpose the two aligned structures and make a note of the RMSD value.

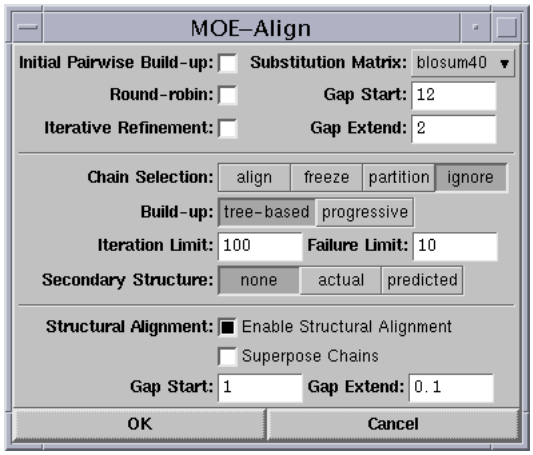


Figure 5: Structural alignment in MOE.

6.3 Repeat sequence and structure alignments for more divergent Pair 2

- Delete the currently loaded proteins. Select chain 1 and chain 2 in the sequence window by shift-clicking the square buttons at the left of each sequence. Select Edit:Delete selected chains
- Repeat, for Pair 2 (`d1eqrb3.ent` and `d1adjc2.ent`, the same steps for loading and performing sequence and structural alignments listed above in section 6.1 and 6.2.

Questions. Describe the relative quality of sequence and structural alignments for Pair 1 and Pair 2. For your assessment use the observed gap lengths and the measured RMSD values. What is the minimum sequence identity between two sequences which allows one to state whether the sequences are homologous (descended from a common ancestor)? To make a phylogenetic tree, which alignment method is better: sequence-based or structure-based?

7 Viewing conserved domains of AARSs

In this section, you will study structural alignments between several different Class II tRNA synthetase catalytic domains, from various organisms.

The required files have been already prepared by P. O'Donoghue[4] and superposed with the structural alignment program STAMP [9]. You will investigate a subset of the Class II tRNA synthetase structures with the program VMD.

- Move to the prepared directory of AARS structures with `cd ~/tbss.work/Bioinformatics/AARS`
- Start VMD by typing at the Unix command prompt: `vmd`
- Open the TkCon console by selecting the VMD menu command Extentions:TkCon
- Now run a TCL script that will, in one step, load all the structures into VMD. Type at the VMD TkCon console:
  - `source loader.tcl`
  - `load_pdb_files .`
- With the TCL script `applyAll.tcl` you can apply different coloring and rendering modes for all loaded structures at once. Try different combinations of rendering and coloring to examine similarities and differences of the structures. Type at the VMD console:
  - `source applyAll.tcl`
  - `applyAll lines backbone`
  - `applyAll cartoon molecule`
  - `applyAll newRibbons structure`
- Try a script that assigns renderings to different parts of the structures. Type at the VMD console:
  - `source applyStructs.tcl`

Questions. What are conserved and variable domains in class II tRNA synthetases? Can you identify these domains in your sequence and structural alignments?

8 Molecular phylogenetic tree.

In this section you will plot a dendrogram displaying the measured similarities between the seven proteins which you pairwise aligned in Section 5. You will compare there relative position in the dendrogram to their relative position in the phylogenetic tree of life.

Below is the pairwise alignment scores from the 21 pairs aligned in section 5. The information in ~/tbss.work/Bioinformatics/multipleData/stats is assembled into a symmetric matrix:

	1EQR	1ATI	1ADJ	1EFW	1ASZ	1B8A	1BBW
1EQR	0.0	0.21	0.23	0.55	0.28	0.31	0.31
1ATI	0.21	0.0	0.25	0.24	0.24	0.18	0.21
1ADJ	0.23	0.25	0.0	0.24	0.24	0.21	0.23
1EFW	0.55	0.24	0.24	0.0	0.34	0.36	0.30
1ASZ	0.28	0.24	0.24	0.34	0.0	0.41	0.27
1B8A	0.31	0.18	0.21	0.36	0.41	0.0	0.28
1BBW	0.31	0.21	0.23	0.30	0.27	0.28	0.0

The commands in the following Matlab session are all in the Matlab script `Dendro.m`. The commands can be run all at once simply by typing `Dendro` at the Matlab command line when Matlab's current directory contains `Dendro.m`.

Now we will use the clustering algorithms in the Statistics toolbox of Matlab to draw a dendrogram of the relatedness of the domains. Here we use the above scores derived from sequence alignment, but structure alignment scores could be used as well[4].

- Move to the directory for this exercise with `cd ~/tbss.work/Bioinformatics/matlabData`
- Start Matlab by typing at the UNIX console: `matlab`.
- The commands in the following Matlab session are all in the Matlab script `Dendro.m`. The commands can be run all at once simply by typing `Dendro` at the Matlab command line, as long as Matlab's current directory contains `Dendro.m and distM.dat`. If you like, type in the below, or paste lines into the Matlab command line from `Dendro.m` or the web-based version of this tutorial. (To see the numerical result of a calculation, leave of the semicolon from the end of the line. To see the value of a variable, enter its name alone on the Matlab command line.)
- First, we read in the above distance matrix of sequence similarity for 7 proteins.

  ```
  load distM.dat;
  ```

- We make a new matrix by subtracting the sequence similarity values from 1, so that longer distances in our dendrogram will correspond to greater evolutionary distance.

```
dM=1-distM;
```

- Its important to keep track of names of the proteins...

```
l={'1eqr','1ati','1adj','1efw','1asz','1b8a','1bbw'};
```

- To use the 'linkage' command of Matlab, one needs to form a column vector of the $((n)(n-1)/2)$ non-redundant elements above the main diagonal of the $n \times n$ distance matrix; our 7×7 matrix produces a 21-element vector:

```
d=[dM(2:7,1);dM(3:7,2);dM(4:7,3);dM(5:7,4);dM(6:7,5);dM(7:7,6)];
```

- Use the `linkage` command to make a hierarchical cluster tree using average distance between cluster elements:

```
z1=linkage(d','average');
```

- For more options in constructing the cluster tree, type `help linkage` at the Matlab command line, also see a modeling text such as Leach [10].
- We display the dendrogram of the clusters in `z1`:

```
h101=figure(101);
dendrogram(z1);
```

- And, finally, paste in some magic to place the labels correctly:

```
hx=get(get(h101,'CurrentAxes'),'XTickLabel');
for i=1:size(hx,1)
   hx(i)=str2double(hx(i));
end
set(get(h101,'CurrentAxes'),'XTickLabel',[l(hx(1)), ...
         l(hx(2)),l(hx(3)),l(hx(4)),l(hx(5)),l(hx(6)),l(hx(7))])
figure(h101);
title('Molecular Phylogenetic Tree');
xlabel('Protein (pdb code)')
ylabel('1-Similarity (%)')
```

- Print out the dendrogram, or copy it down on paper, along with the names of the proteins. Refer to Table 6 to write, under each name, the domain of life each protein originates from.

Questions. What is the pair with the closest evolutionary relation? What is the pair with the most distant relation? Is the arrangement of the proteins in the dendrogram consistent with what we know about the evolution of the three domains of life?

9 Other bioinformatics tools

So far in this tutorial, you have made use of only a small selection of bioinformatics techniques and tools. In the last exercise we invite you to explore additional tools and resources by yourself. Results of aligning sequences can be improved by systematically building up profiles from multiple sequences.

- Try using the multisequence alignment servers such as ClustalW or servers employing Hidden Markov methods to build a profile from the four aspartyl AARSs sequences.
- Align the histidyl AARS to the profile.
- Check if you can obtain an alignment closer than the structure-based alignment you saw in MOE.

Tools, resources, and link collections:

ClustalW (`http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html`)
Perform a multi-sequence or profile-profile alignment with the program ClustalW. Just access the website directly and paste in all or a selection of your Class II tRNA synthetases in order to execute the program. ClustalW is the most widely used tool in bioinformatics for carrying out multi-sequence alignments.

Psipred (`http://bioinf.cs.ucl.ac.uk/psiform.html`)
Predict the secondary structure of one of your Class II tRNA synthetases with the Psipred Protein Structure Prediction Server. Paste your sequence in the input sequence window, provide your email address and you will receive after a few minutes a secondary structure prediction of your chosen tRNA synthetase. Sequence and structural alignments as well as secondary predictions form the framework for a successful modeling project.

3D PSSM (`http://www.sbg.bio.ic.ac.uk/~3dpssm/`)
A web-based method method for protein fold recognition using sequence profiles coupled with secondary structure.

TMpred (`http://www.ch.embnet.org/software/TMPRED_form.html`)
A database scoring-based method to predict the transmembrane portions of membrane proteins.

TMHMM (`http://www.cbs.dtu.dk/services/TMHMM-2.0/`)
A hidden Markov method to predict the transmembrane portions of membrane proteins.

European Bioinformatics Inst. (`http://www.ebi.ac.uk/services/index.html`)
An up-to-date and well-organized collection of links to bioinformatics tools, databases, and resources. The site provides advice as to the best or most popular tools in a category, and provides short descriptions of all entries.

ExPASy Molecular Biology Server (`http://ca.expasy.org/`)
Another well-organized directory of online analysis tools, databases, and other resources, with a greater focus on proteins. "The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures..." With this

server you can start your own homology modeling project of an unknown class II tRNA synthetase, namely Alanyl-tRNA synthetase. You can obtain the sequence in FASTA format from the SwissProt database which can be accessed directly from the ExPASy server with the accession number SYA_ECOLI. As structural template choose one of the provided catalytic domain structures of class II tRNA synthetases. You can also model the other domains for which you need to find an appropriate template from the provided PDB structures.

SwissModel (`http://swissmodel.expasy.org/`)
For model generation use SwissModel, where you can thread your sequence upon one or several of your chosen templates. SwissModel provides you with an online tutorial and will perform refinements on initial models you submit to its server.

Dynamic Programming in Java
(`http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html`)
This is an alternative Smith-Waterman tutorial which will provide you with a web-based interface for dynamic programming, an animated version of the paper-and-pencil exercise in section 5.

Biology WorkBench (`http://workbench.sdsc.edu`)
This website allows you to search popular protein and nucleic acid sequence databases. Sequence retrieval is integrated with access to a variety of analysis tools as for example the multi-sequence alignment program ClustalW. The advantage of the Biology Workbench is that all analysis tools are interconnected with each other eliminating the tedious file conversion process, which often needs to be done when accessing tools from distinct locations.

CASP5 (`http://predictioncenter.llnl.gov/casp5/Casp5.html`)
Every two years a community-wide protein structure prediction contest takes place,where groups complete for prediction of unpublished protein structures. One can check out how well has our Resource done in the last year contest. Just search for Zan Schulten Group results on this site.

References

[1] Michael S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes*. CRC Press, 1995.

[2] M. Ibba and D. Söll. Aninoacyl-tRNA synthesis. *Ann. Rev. Biochem*, 69:617–650, 2000.

[3] C. Woese, G. Olsen, M. Ibba, and D. Söll. Aminoacyl-RNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Bio. Rev.*, pages 202–236, 2000.

[4] P. O'Donoghue and Z. Luthey-Schulten. On the evolution of structure in the aminocyl-tRNA synthetases. 2003. In preparation.

[5] C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273:1058–1073, 1996.

[6] S. E. Benner, P. Koehl, and M. Levitt. The ASTRAL compendium for seqeunce and structure analysis. *Nucleic Acid Research*, pages 254–256, 2000.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[8] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[9] R. B. Russel and G.J. Barton. Multiple protein sequence alignment from tertiary struture comparison: assignment of global and resiude confidence levels. *Proteins: Struct., Func., Gen.*, 14:309–323, 1992.

[10] Andrew R. Leach. *Molecular Modelling: Principles and Applications (2nd edition)*. PrenticeHall, Upper Saddle River, New Jersey, 2001.